UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Modelo de Propensão à Conversão

Ana Catarina Piedade Miranda

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado por:
Professora Doutora Maria Isabel Fraga Alves

2019

“Bom mesmo é ir à luta com determinação,
abraçar a vida com paixão,
perder com classe e vencer com ousadia,
porque o mundo pertence a quem se atreve.
E a vida é muito bela para ser insignificante.”

Charles Chaplin

Agradecimentos

Aos meus pais, por todo o apoio e por sempre me terem apoiado e incentivado a progredir com os meus estudos.

À minha orientadora, Professora Maria Isabel Fraga Alves, por todo o carinho e paciência disponibilizados.

Aos meus colegas da AGEAS e a toda a equipa do PBA por toda a ajuda despendida e por todo o companheirismo e interesse ao longo do estágio.

À AGEAS pela oportunidade da realização do estágio, que me deu o primeiro impulso no mundo do trabalho e me fez aprender e crescer bastante.

E por fim, a todos os meus amigos e família pelo apoio e motivação que me deram para acabar este projeto.

Resumo

O aumento do mercado segurador tem levado a que exista uma maior competitividade entre as entidades seguradoras, tendo estas como principal foco o desenvolvimento de novas abordagens para angariarem cada vez mais clientes, e conquistarem uma posição líder no mercado. É por isso cada vez mais importante para as seguradoras conhecerem os seus clientes, para que possam desenvolver medidas de forma a angariar aqueles que são considerados pelas seguradoras os 'bons' clientes, aqueles que lhes possam proporcionar um maior lucro no futuro.

Este relatório é o resultado de um estágio profissional que teve lugar na AGEAS Seguros Portugal, na equipa de *Pricing and Business Analytics*, inserida na direção Técnica e Oferta Não Vida. É nesta área de negócio que são desenvolvidas e implementadas estratégias de ajuste aos preços dos produtos que posteriormente serão apresentados aos clientes.

Este estágio teve como objetivo o desenvolvimento de um modelo de propensão à conversão, para o setor automóvel da companhia e identificação dos perfis de clientes com maior e menor propensão á conversão, delineando ações para cada um destes segmentos. Este projeto tem como principais benefícios para a empresa, potenciar as vendas e a rentabilidade e uma melhor compreensão do comportamento dos clientes.

Para a construção da base de dados foram utlizadas as simulações relativas a uma nova tarifa para o sector automóvel, que foi implementada em setembro de 2017, sendo assim o período temporal da base de dados em estudo de 26 de setembro de 2017 a 14 de janeiro de 2019. Estes serão os dados que refletem o perfil de potencial cliente relativamente à conversão, e face à oferta da nova tarifa automóvel.

Foi utilizada a regressão logística, um caso particular dos Modelos Lineares Generalizados, para modelar a variável binária de resposta, associada às ações "converteu/não converteu". Constituiu o nosso objetivo a modelação da variável resposta em função das restantes variáveis explicativas pertinentes para o estudo.

Para o desenvolvimento deste projeto foi necessário realizar uma cuidadosa preparação da base de dados, foram realizadas análises descritivas das variáveis explicativas que poderão fazer parte do modelo e foram também, explorados comparativamente vários modelos recorrendo a dois métodos distintos, árvores de decisão e regressão logística; por fim, e após a análise de índices de seleção convenientes, será escolhido o melhor modelo, ou seja, o que se revela mais adequado às necessidades da companhia, mediante o leque de modelos estudados.

Palavras-chave: Seguro Automóvel, Taxa de Conversão, Modelo de Regressão Logística

Abstract

The growth of the insurance market has led to greater competitiveness among insurance providers, with the main focus being on developing new approaches to attract more and more customers, and to gain a leading position in the market. It is therefore becoming more important for insurers to know their clients so that they can develop measures to attract those who they consider to be 'good' clients, those who can provide them with greater profit in the future.

This report is the result of a professional internship in AGEAS Seguros Portugal, in the *Pricing and Business Analytics* team, in the Technical and Non-Life Offering department. It is in this business area that strategies are developed and implemented to adjust the prices of products that will later be presented to customers.

This internship aimed to develop a conversion propensity model for the automobile sector of the company and identification of client profiles with a major and minor conversion propensity, outlining actions for each of these segments. This project has, as the main benefits for the company, boosting sales and profitability and a better understanding of client behaviour.

For the construction of the database, simulations of a new automobile section tariff were used, which was implemented in September 2017; therefore the time period of the database under study is from 26 September 2017 to 14 January. 2019. This will be the data that reflects the potential client profile relative to the conversion and in light of the offer of the new motor tariff.

Logistic regression, a particular case of Generalized Linear Models, was used to model the binary response variable associated with “converted / not converted” actions. Our objective was the modeling of the response variable as a function of the remaining explanatory variables relevant to the study.

For the development of this project it a careful preparation of the database was required, followed by analysis of explicative variables which could be part of the model and also comparatively explored several models using two different methods, decision trees and logistic regression; finally, and after the analysis of suitable selection indices, the best model will be chosen, that is, that which is more suitable to the needs of the company, within the range of models studied.

Keywords: Motor Insurance, Conversion Rate, Logistic Regression Model

Índice

Lista de Gráficos

Lista de Tabelas

Lista de Figuras

Glossário

AGE – Agente exclusivo à companhia, apenas trabalham com a AGEAS;

Agente – Qualquer pessoa ou entidade que exerça a atividade de mediação de seguros, e se encontre inscrito como mediador na Autoridade de Supervisão de Seguros e Fundos de Pensões. Pode fazê-lo por conta de um ou vários seguradores ou de forma independente;

Agravamento – Aumento do prémio na renovação do contrato de seguro, após ser verificada a existência de sinistros no período de contrato antes da renovação.

Apólice de Seguro – Documento que titula o contrato celebrado entre o tomador do seguro e a empresa de seguros, de onde constam as respetivas condições gerais, especiais, se as houver, e particulares acordadas;

ASF – Autoridade de Supervisão de Seguros e Fundos de Pensões, entidade responsável pela regulação e supervisão da atividade seguradora, resseguradora, dos fundos de pensões e das entidades gestoras e mediação de seguros;

Bonificação – Redução do prémio na renovação do contrato de seguro, após ser verificada a ausência de sinistros no período de contrato antes da renovação;

Bónus – Redução do prémio de renovação do contrato de seguro, após serem verificadas determinadas circunstâncias fixadas na apólice, com por exemplo a ausência de sinistros;

CAMEO – Base de dados externa à companhia que contém a informação relativa ao código postal;

Capital Seguro – Montante estipulado nas condições particulares do contrato como sendo o limite máximo de responsabilidade da empresa de seguros;

Carteira – Conjunto de contratos de seguro ou dos contratos de capitalização subscritos junto de uma empresa de seguros;

Cliente Não Novo – Cliente que já possuiu apólice automóvel na companhia;

Companhia de Seguros – Entidade legalmente autorizada a exercer a atividade seguradora e que é parte no contrato seguro;

Contrato de Seguro – Operação comercial pela qual uma parte, a empresa de seguros, se compromete, mediante o recebimento de um pagamento, e na eventualidade de ocorrer um evento aleatório, a fornecer à outra parte contratante uma prestação em dinheiro ou serviço;

Conversão – Quando uma simulação se transforma em apólice, ou seja, quando o cliente adquire um seguro na companhia;

Franquia – Dano ou parte do dano que fica a cargo do segurado;

Prémio – O prémio bruto acrescido das cargas fiscais e parafiscais, e que corresponde ao preço pago pelo tomador de seguro à empresa de seguros pela contratação do seguro;

Prémio Comercial – Custo teórico médio das coberturas do contrato, acrescido de outros custos, nomeadamente a aquisição e de administração do contrato, bem como de gestão e de cobrança;

Seguro de Danos Próprios – Seguro que garante a reparação ou a substituição de um veículo terrestre após choque, colisão, capotamento, incêndio, raio ou explosão e furto ou roubo;

Seguro de Responsabilidade Civil – Seguro que garante as consequências pecuniárias da responsabilidade que compete ao segurado, em consequência de danos causados a outrem e provocados pelo próprio segurado, por pessoas por quem ele é responsável ou por animais ou bens que tem à sua guarda;

Seguro de Responsabilidade Civil Automóvel – Seguro de responsabilidade civil que cobre os danos causados a terceiros por veículos terrestres a motor e seus reboques, sendo um seguro obrigatório;

Sinistro – Evento ou série de eventos resultantes de uma mesma causa capaz de fazer funcionar as garantias de um ou mais contratos de seguro;

Tarifa – Designação dada ao quadro de prémios ou taxas de prémio a aplicar aos riscos a segurar e ao conjunto de condições de subscrição de um dado ramo;

Terceiro – A vítima de um sinistro que não é parte no contrato de seguro mas que, por força deste, assume o direito a ser indemnizada nos termos do mesmo;

Tomador do Seguro – Pessoa singular ou coletiva que celebra o contrato de seguro com a empresa de seguros, sendo responsável pelo pagamento do prémio;

Fonte: ASF – Autoridade de Supervisão de Seguros e Fundos de Pensões

Introdução

Hoje em dia, os clientes procuram obter sempre a melhor qualidade/preço antes de adquirir um seguro, tanto automóvel como nas outras áreas. Daí a competitividade entre as seguradoras ser cada vez mais evidente.

O conhecimento do cliente, ou seja, os seus gostos e necessidades e principalmente o seu comportamento, tem sido uma mais-valia para as companhias seguradoras, pois assim conseguem salvaguardar parte do seu investimento, evitando os 'maus' clientes e procurando conquistar aqueles que são considerados no universo segurador como os 'bons' clientes. Os 'bons' clientes são aqueles que apresentam taxas de sinistralidade bastante reduzidas, permitindo assim às companhias retirar lucro com os seus seguros.

É através das simulações, que tanto os clientes como as companhias seguradoras conseguem retirar uma grande mais-valia, os clientes porque utilizam este método para estudar o mercado e poder encontrar o seu seguro ideal no que diz respeito ao fator qualidade/preço, já as companhias seguradoras conseguem reter informações dos clientes para o desenvolvimento de modelos e estudar formas de adquirir mais clientes.

O seguro automóvel é um dos seguros obrigatórios por lei, obrigando todos os condutores a possuírem pelo menos a cobertura mais simples, de responsabilidade civil, tendo como principal objetivo a proteção do condutor e de eventuais vítimas que possam surgir caso exista um sinistro.

Como já foi anteriormente referido, o objetivo deste projeto é modelar a taxa de conversão de clientes automóvel para a seguradora em questão. Para tal foram utilizados dois métodos que nos permitem criar modelos preditivos, as árvores de decisão e a regressão logística, utilizando uma variável resposta de caracter binário, em que toma o valor 0 se o cliente não converter em apólice e o valor 1 caso o cliente converta em apólice, ou seja, passe a ter uma apólice de seguro na companhia. Serão posteriormente analisados estes dois métodos de modo a escolher o que apresenta um modelo melhor.

Para a preparação dos dados e desenvolvimento dos modelos serão utilizadas duas ferramentas do SAS, o SAS Guide para a construção da base de dados e para a análise descritiva das variáveis e o SAS Miner para a criação dos modelos. Ambas as ferramentas de análise foram disponibilizadas pela companhia.

Relativamente aos dados em estudo, estes dizem respeito a uma nova tarifa desenvolvida pela companhia, AGEAS, esta teve início em setembro de 2017 e o estudo irá até meados de janeiro de 2019. Como podemos observar pelo gráfico seguinte, onde estão representadas as simulações e as conversões realizadas pelos clientes durante esse período. É nos fechos do ano, ou seja, no mês de dezembro que se registam as taxas de conversão mais elevadas, embora os meses de verão também apresentem uma taxa de conversão elevada.

Para a realização deste projeto serão utilizadas árvores de decisão e modelos lineares generalizados, recorrendo a um caso particular destes, a regressão logística. Todos os modelos serão posteriormente avaliados para que se possa escolher o melhor modelo para a companhia.

No gráfico seguinte, podemos observar as tendências relativas á taxa de conversão automóvel no período em estudo para a realização do modelo. É nos últimos meses do ano que se observam taxas de conversão mais elevadas, como podemos observar através da linha apresentada no gráfico.

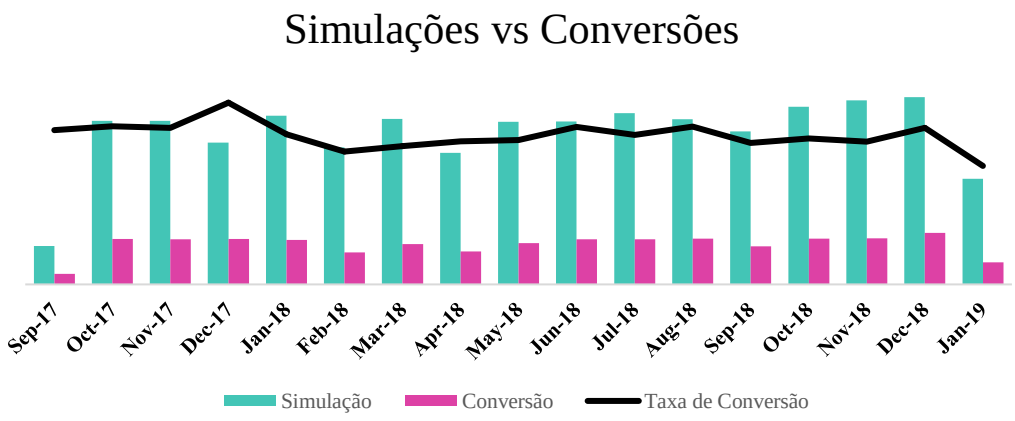

Gráfico 1 – Sazonalidade das Simulações e Conversões

Capítulo 1: Atividade Seguradora

1.1 História da Atividade Seguradora

É na idade Média, que se abrem as portas para o seguro moderno, quando pela primeira vez existe a transferência de risco para um terceiro. Esta transferência ocorre com um contrato de empréstimo. Para armarem os barcos em que eram transportadas as mercadorias, os comerciantes precisavam de fundos, recorrendo a empréstimos, que seriam pagos com um juro até 50% caso a viagem fosse bem-sucedida, mas caso contrário o empréstimo não seria reembolsado. Mais tarde esta prática acabou por ser proibida.

A proibição criou a necessidade de outras soluções, no qual se chegou a um esquema em que a operação de garantia era separada do empréstimo, ou seja, os comerciantes associavam-se e aceitavam garantir, mediante um pagamento previamente fixado, para que caso houvesse um naufrágio, quem perdesse o barco e a respetiva carga seria reembolsado.

A importância económica de Génova e Veneza, devido ao comércio marítimo e a Florença devido ao sistema bancário, faz com que estas cidades venham, nos séculos XIV a XVII, a construir os principais centros da atividade seguradora na Europa.

O seguro de vida, só aparece mais tarde, no final do século XVII, por estudos realizados por Pascal e por Halley.

Mais tarde, em 1666, em Londres, ocorreu um catastrófico incêndio, fazendo milhares de vítimas. Esse incêndio levou a que fosse criado o primeiro corpo de bombeiros na Europa, e em 1684, ao aparecimento do seguro contra o risco de incêndio.

Relativamente a Portugal, foi entre 1293 e 1297, no reinado de D. Dinis, que com o desenvolvimento do comércio marítimo, que se começa a afirmar a atividade seguradora nacional. No reinado de D. Fernando foi criado uma espécie de mercado segurador, a “Bolsa de Seguros”.

Podemos então constatar que o seguro surge da sobreposição de três elementos:

- Existência de riscos comparáveis e capazes de se compensarem entre si;
- Componente jurídica, que se materializa no contrato de seguro;
- Componente de conhecimento científico: a provisão do risco, é realizada com base na informação estatística passada relativa à ocorrência. (Guedes-Vieira, 2012)

1.2 AGEAS

1.2.1 AGEAS no Mundo

A AGEAS é um grupo segurador internacional, com sede em Bruxelas e com 190 anos de experiência e de conhecimento. Atualmente está presente em 14 países da Europa e da Ásia, a empresa propõe soluções vida e não vida a milhões de clientes individuais e empresas.

Reconhecida pela sua forte experiência em matéria de parcerias, a AGEAS desenvolve acordos de longa duração com instituições financeiras e distribuidores locais de referência pelo mundo inteiro, de forma a garantir a proximidade com os seus clientes.

A AGEAS é um dos maiores grupos seguradores, é líder na Bélgica e encontra-se entre os principais *players* na maioria dos países em que está presente. Tem mais de 40.000 colaboradores e encontra-se presente na Bélgica, Reino Unido, França, Portugal, Turquia, China, Malásia, India, Tailândia, Vietname, Laos, Camboja, Singapura e Filipinas.

1.2.2 AGEAS em Portugal

O grupo AGEAS Portugal é um dos líderes no ranking segurador português, operando em Portugal desde 2005 através de marcas conhecidas como a Ocidental e a *Médis*, tendo-se juntado a AGEAS Seguros e a Seguro Direto em 2016.

A AGEAS aposta em Portugal como um dos principais mercados, onde se pretende desenvolver, através de parcerias fortes e contribuindo para o desenvolvimento do país e da sociedade, ajudando os clientes a gerir, antecipar e proteger-se contra riscos e imprevistos, para que possam viver o presente e o futuro com a máxima segurança e serenidade.

1. 3 Mercado da Atividade Seguradora em Portugal

As empresas de seguros têm vindo a oferecer uma grande diversidade de produtos e serviços tendo em consideração não só a proteção do risco da vida humana e dos seus bens, mas também as poupanças e os rendimentos de capital.

O mercado segurador tem vindo a crescer em Portugal ao longo dos anos, e este demonstra um papel fundamental na economia nacional.

O setor segurador representa, atualmente, cerca de aproximadamente 7% do PIB. É no ramo vida que se tem registado um maior crescimento de cerca de 7%, devido à melhoria das condições macroeconómicas com impacto positivo na atividade económica do país.

As modalidades que mais sobressaem atualmente, tanto no ramo vida como no ramo de não vida, por serem mais procurados pelos portugueses dizem respeito, aos seguros automóvel e acidentes de trabalho, ao seja, aos seguros obrigatórios.

1.4 Tipos de Seguros

No ramo segurador existe uma grande diversidade de seguros disponíveis no mercado, dividindo-se em duas modalidades:

1.4.1 Ramo Vida

Este ramo inclui os seguintes seguros e operações:

- Seguro de Vida – Seguro efetuado sobre a vida de uma ou mais pessoas e garante principalmente, o risco de morte ou o risco de sobrevivência;
- Seguro de Nupcialidade – Pagamento de um prémio ou de uma renda caso a pessoa segura se case;
- Seguro de Natalidade – Pagamento de um prémio ou de uma renda em caso de nascimento de filhos;
- Seguro de Fundos de Investimento Coletivo – Seguro em que as importâncias seguras são determinadas em função de um "valor de referência constituído por uma "unidade de conta" ou pela combinação de várias unidades de conta;

1.4.2 Ramo Não Vida

No ramo não vida existem vários seguros, sendo os mais usuais:

- Automóvel – Seguro que se divide em duas componentes, Responsabilidade Civil e Danos Próprios, sendo a primeira componente de caracter obrigatório.
- Acidentes de Trabalho – Seguro obrigatório, que visa a proteção do indivíduo contra qualquer acidente que ocorra no seu local de trabalho ou no trajeto do mesmo;
- Acidentes Pessoais – Seguro que visa a proteção do indivíduo contra qualquer acidente que ocorra no decurso do seu dia e que esteja fora do âmbito dos seguros obrigatórios;
- Saúde - Seguro que cobre riscos relacionados com a prestação de cuidados de saúde, conforme as coberturas previstas nas condições do contrato;

- Multirriscos – Seguro que cobre os riscos relacionados com a habitação. Consoante as coberturas que o cliente subscrever, pode garantir, que caso ocorra algum imprevisto, a sua habitação e os bens que esta contenha ficarão segurados.

Capítulo 2: Seguro Automóvel

2.1 Seguro Automóvel

É no ramo automóvel que está representada a maior parte das carteiras das seguradoras portuguesas. No entanto é neste ramo que se verificam as taxas de sinistralidade mais elevadas.

O seguro automóvel divide-se em dois grupos, o seguro de responsabilidade civil, que é um seguro obrigatório e o seguro que danos próprios que é um seguro facultativo.

2.1.1 Seguro de Responsabilidade Civil

O seguro automóvel é um seguro obrigatório no que diz respeito à cobertura de responsabilidade civil, que tem como função a proteção das vítimas, caso exista um sinistro.

Um veículo para o qual não foi contratado seguro de responsabilidade civil encontra-se numa situação ilegal. Por lei, o veículo pode ser apreendido e o seu proprietário pode ter de pagar uma coima.

Este seguro dá resposta ao que é legalmente exigido no que diz respeito a qualquer veículo terrestre a motor, reboques e semi-reboques. Esta cobertura garante que em caso de danos provocados a terceiros, sejam estes corporais ou materiais, que possam ser devidamente indemnizados pelas seguradoras.

As pessoas seguras por este seguro são: peões da via pública ou que circulem noutra viatura, que circulem em transportes coletivos de passageiros ou a título gratuito.

O seguro obrigatório assegura o pagamento das indemnizações por danos corporais e materiais causados a terceiros e às pessoas transportadas, com exceção do condutor do veículo. No mínimo, este seguro tem de cobrir cinco milhões de euros para danos corporais e um milhão de euros por acidente para danos materiais.

O seguro de responsabilidade civil encontra-se válido no território nacional, nos restantes estados membros da União Europeia e em países exteriores à União Europeia que tenham aderido ao acordo multilateral de garantia entre serviços nacionais de seguros.

2.1.2 Seguro de Danos Próprios

O Seguro de danos próprios é um seguro facultativo que colmata algumas coberturas excluídas no seguro de responsabilidade civil. Este tipo de seguro cobre todos os prejuízos sofridos pelo veículo seguro independentemente de quem seja o responsável pelo sinistro.

Este seguro possui garantias base que estão limitadas apenas ao território nacional, essas garantias são: choque, colisão e capotamento; quebra isolada de vidros; incêndio, raio e explosão; furto ou roubo. Existem também outras garantias que poderão ser adquiridas, como é o caso dos fenómenos da natureza, atos de vandalismos, assistência em viagem, entre outros.

2.2 Tarifa

A tarifa consiste num conjunto de regras que permitem definir o prémio a pagar por cada apólice. Esta surge como instrumento que contempla os prémios, bem como as regras orientadoras para a realização de contratos de seguros de um determinado ramo.

No ramo Automóvel, as tarifas são calculadas tendo em consideração as características do tomador de seguro e do veículo seguro, bem como alguma informação que seja considerada relevante, como por exemplo: localidade do segurado, de forma a que se consiga obter um valor adequado para o prémio a ser pago pelo cliente.

2.3 Sistema de Bónus *Malus*

O sistema de bónus *malus* apenas é aplicável aos seguros automóvel, sendo apenas afetadas as seguintes coberturas: responsabilidade civil, choque, colisão ou capotamento, incêndio, raio ou explosão, furto ou roubo e só colisão.

Este sistema é a forma que as companhias têm de beneficiarem ou agravarem os prémios dos clientes consoante o seu histórico de sinistralidade.

A percentagem de bónus *malus* a ser aplicada a cada cliente difere de companhia para companhia, e esta atribuição depende do número de anos em que o contrato está em vigor e do número de sinistros registados, isto é, o histórico comprovado do segurado a partir do certificado de tarifação, no caso de transferências externas, ou do histórico da companhia, por transferências internas, e não da percentagem de bonificação ou agravamento no contrato anterior. A determinação do escalão inicial do bónus *malus* tem em consideração o número de sinistros nos últimos 5 anos e os anos com seguro dos clientes.

Por exemplo, se um determinado cliente apresenta um escalão entre 0 e 6 diz-se que este cliente apresenta um malus e será aplicado um agravamento ao seu prémio. No entanto, se o escalão estiver entre 7 a 26, o cliente apresenta um bónus, onde irá ter uma bonificação sobre o seu prémio. Sempre que um cliente tenha um sinistro, o seu escalão é revisto, descendo assim de escalão de bónus.

2.4 Packs Automóvel

Na tabela estão representados os Packs automóvel que a AGEAS oferece.

Coberturas	Pack 1	Pack Extra	Pack Confort	Pack 4
Responsabilidade civil obrigatória Danos Materiais €1.000.000 Danos Corporais €5.000.000	•	•	•	•
Responsabilidade Civil €50.000.000	• (opcional)	• (opcional)	• (opcional)	• (opcional)
Assistência em Viagem	• **Normal** Vip (opcional)	• **Normal** Vip (opcional)	• **Normal** Vip (opcional)	• Vip (opcional)
Proteção Jurídica	•	•	•	•
Proteção Ocupantes Morte ou Invalidez Permanente Despesas de Tratamento Despesas de Funeral Incapacidade Temporária Absoluta por Internamento Hospitalar	• **Só Condutor** Todos os Ocupantes (Opcional)	• **Só Condutor** Todos os Ocupantes (Opcional)	• **Só Condutor** Todos os Ocupantes (Opcional)	• **Só Condutor** Todos os Ocupantes (Opcional)
Quebra de Vidros	**(Opcional)** Com Franquia	• Com Franquia	• Com Franquia	• Com Franquia
Choque, Colisão ou Capotamento			•	•
Incêndio, Raio ou Explosão		•	•	•
Furto ou Roubo		•	•	•
Fenómenos da Natureza (FN) e Riscos Sociais (RS)			**(Opcional)** Com Franquia em RS	•
Veículo Aluguer			**(Opcional)**	•

Veículo Novo				•
Solução Ano Seguro		**(Opcional)**	**(Opcional)**	•
Bagagem Pessoal				•
Seguro de Pneu	**(Opcional)**	**(Opcional)**	**(Opcional)**	**(Opcional)**

Tabela 1 - Descrição dos Packs

Capítulo 3: Enquadramento Teórico

3.1 Árvores de Decisão

Podemos caracterizar as árvores de decisão como ferramentas de classificação e predição de observações com base num conjunto de regras de decisão. As árvores de decisão são compostas por nós, arcos e folhas. Os nós, representam as interrogações que se fazem sobre um conjunto de dados. Os arcos ou resultados, separam o conjunto de dados de acordo com a resposta à interrogação. Finalmente, as folhas representam os nós finais onde já não existe qualquer interrogação.

A classificação será tanto melhor, quanto melhor for a qualidade dos dados. As regiões definidas pelas folhas são mutuamente exclusivas, ou seja, a interseção de regiões abrangida por quaisquer duas folhas é vazia. As árvores de decisão são ferramentas capazes de construir bons modelos preditivos, sendo particularmente relevantes quando a variável de *output* é uma variável discreta e assume um número reduzido de valores. Possuem também a capacidade de utilização de dados binários e categóricos sem qualquer necessidade de transformação.

No exemplo seguinte, *figura 1*, está representada uma árvore de decisão que codifica os diferentes percursos de avaliação de alunos de um determinado curso. Assim, os alunos caso tenham uma nota igual ou superior a 10 valores são aprovados, caso contrário, se tiverem uma nota no trabalho inferior a 14 reprovam, se tiverem uma nota no trabalho igual ou superior a 14 e uma nota no exame igual ou superior a 8 valores são aprovados.

A *figura 1* permite-nos também observar de uma forma mais simples a constituição das árvores de decisão.

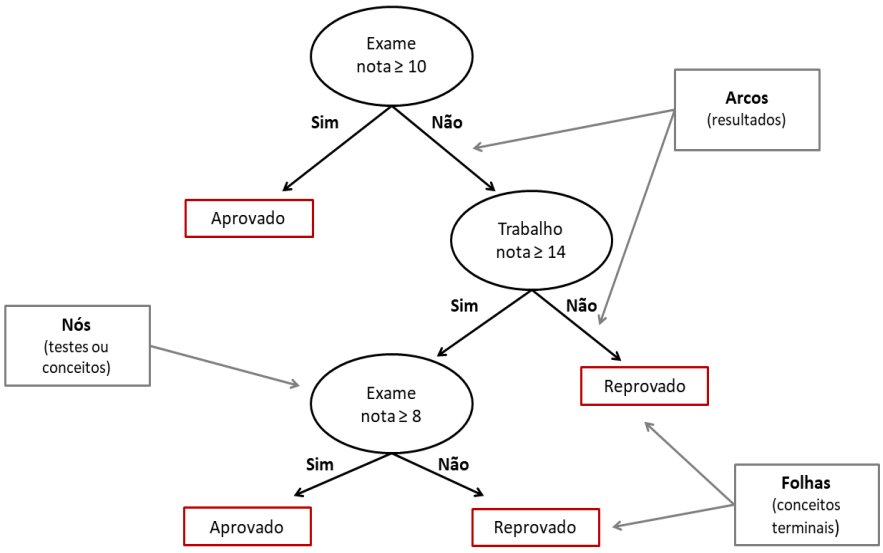

Figura 1 - Exemplo Árvore de Decisão

3.1.1 Indução pelo Algoritmo CART

O algoritmo CART (*Classification And Regression Trees*) é um dos métodos mais utilizados para a construção de árvores de decisão. Este algoritmo constrói uma árvore binária, onde cada nó apenas dá origem a dois nós descendentes. O CART separa os registos em cada nó de acordo com uma função de um único campo de *input*. Assim, a primeira tarefa consiste em decidir qual, de entre as variáveis de *input*, produz a melhor partição. A melhor partição é definida como sendo a que produz a melhor separação dos registos em grupos onde uma única classe predomina.

Para escolher a melhor partição num determinado nó, são consideradas as variáveis de *input*, uma de cada vez. Como as variáveis possuem imensos valores, em primeiro lugar, estes registos são ordenados com base nas diferentes variáveis. De seguida, é avaliada cada partição possível de acordo com o valor da diversidade (com base no coeficiente de *Gini*). Assim, os valores assumidos pelas observações para as diferentes variáveis de *input* constituem as partições que podemos testar em termos de diversidade. Isto significa que iremos procurar todas as partições possíveis, avalia-las tendo em conta o valor da diversidade e por último, escolher a que apresenta uma diversidade menor.

O processo de crescimento da árvore ocorre até não ser possível criar mais subárvores, ou, alternativamente, até atingir uma condição de paragem previamente definida. É por isso importante a escolha de um critério de paragem. Um dos critérios é o estabelecimento de um limite para a medida de diversidade, ou seja, quando este atinge um valor abaixo do limite estabelecido a árvore para de crescer. Outro critério bastante utilizado passa pela possibilidade de impedir o desenvolvimento da árvore a partir do momento em que os nós possuam um número de registos inferior a um limite estipulado. A lógica subjacente a este critério tem a ver com a representatividade do nó. Quando o nó apresenta um número bastante reduzido de observações é natural que qualquer particularidade daquele pequeno conjunto possa reduzir a diversidade.

3.1.2 Coeficiente de *Gini*

O algoritmo de indução tem que escolher qual o atributo preditivo que será utilizado em cada nó da árvore. Essa escolha será baseada em diferentes critérios, como a impureza, a distância ou a dependência.

O critério utilizado será o coeficiente de *Gini*, que é baseado no grau de pureza do nó. A medida de *Gini* é dada por:

$$Gini(N) = 1 - \sum_{C=1}^{k} p(C|N)^2 \qquad (3.1)$$

em que $p(C|N)$ é a fração de elementos que pertencem à classe *C*, no nó *N*, e o *k* corresponde ao número de classes. Desigualdades maiores entre as diferentes proporções das classes originam um índice de *Gini* mais reduzido, verificando-se o oposto quando as proporções são mais

equilibradas. Assim, a variável originária da melhor partição é aquela que origina uma maior redução do índice de *Gini* nos ramos que dela resultam.

3.2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados são casos específicos de modelos de regressão, que têm como principal objetivo o estudo da relação entre variáveis, ou seja, a análise do efeito que as variáveis explicativas provocam sobre a variável resposta.

Estes modelos apresentam vantagens por serem uma generalização dos modelos lineares, entre as quais, a flexibilidade da função de regressão, ou seja, da relação existente entre a variável resposta e a combinação linear das variáveis explicativas. Esta relação é habitualmente efetuada pela função de ligação. A maior vantagem consiste na possibilidade de construção de modelos que permitem obter intervalos de confiança das estimativas, com base em distribuições pertencentes à família exponencial (Turkman, 2000).

Os Modelos Lineares Generalizados são bastante utilizados na área dos seguros, pois para a análise deste tipo de dados o modelo linear normalmente não é aplicável. Por exemplo, os tamanhos dos sinistros, as frequências dos sinistros e a ocorrência de uma reclamação são dados que usualmente não evidenciam uma distribuição normal (Jong, 2008).

3.2.1 Família Exponencial

Os Modelos Lineares Generalizados partem do princípio que a variável resposta tem uma distribuição pertencente a **Família Exponencial**.

Podemos dizer que uma variável *Y* possui uma distribuição pertencente à família exponencial se a sua função densidade de probabilidade ou a sua função massa de probabilidade se puder escrever da seguinte forma:

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)}\right\} + c(y,\phi) \qquad (3.2)$$

onde, θ é a forma canónica do parâmetro escalar de localização; ϕ é um parâmetro escalar de dispersão conhecido; $a(.), b(.) e\ c(.,.)$ são funções reais conhecidas. (Turkman, 2000).

3.2.2 Regressão Linear Simples

Um modelo de regressão estabelece uma relação de causa-efeito entre duas ou mais variáveis.

O **modelo de regressão linear simples** analisa a relação entre duas variáveis, uma variável resposta (Y) e uma variável explicativa (X), cuja tendência é aproximadamente linear, e é dado pela seguinte expressão:

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}. \quad (3.3)$$

De acordo modelo (3.3) obtém-se:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n}. \quad (3.4)$$

onde, y_i representa o valor da variável resposta, Y, na observação $i = 1, \dots, n$ (aleatória); x_i representa o valor da variável independente, X, na observação $i = 1, \dots, n$ (não aleatória); β_0 e β_1 são parâmetros do coeficiente de regressão, e $\varepsilon_i, i = 1, \dots, n$ são variáveis aleatórias que correspondem ao erro.

Com base nos dados disponíveis, queremos estimar os parâmetros β_0 e β_1, o que é equivalente a encontrar a linha reta que nos dá o melhor ajuste dos pontos no gráfico de dispersão da variável resposta versus a variável preditora. Estimamos os parâmetros usando o método dos mínimos quadrados, que consiste na obtenção dos estimadores dos coeficientes de regressão β_0 e β_1, minimizando a seguinte soma de quadrados:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \quad (3.5)$$

Os valores de $\widehat{\beta_0}$ e $\widehat{\beta_1}$ que minimizam $S(\beta_0, \beta_1)$ são dados por:

$$\widehat{\beta_1} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad \text{e} \quad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}. \quad (3.6 \text{ e } 3.7)$$

Para cada observação dos nossos dados, podemos calcular a reta dos mínimos quadrados que é dada pela seguinte equação:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i, \quad i = 1, \dots, n. \quad (3.8)$$

A equação de regressão ajustada pode ser usada para **predição**. Existem dois tipos de predição:

1. A predição do valor da variável resposta, Y, que corresponde a qualquer valor escolhido, x_0, da variável preditora. Neste caso o valor previsto é dado por:

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0. \quad (3.9)$$

2. A estimativa do valor médio μ_0 quando $X = x_0$, a estimativa é dada por:

$$\widehat{\mu_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0. \quad (3.10)$$

(Chatterjee et al, 2012)

3.2.3 Regressão Linear Múltipla

Na regressão linear múltipla assume-se que existe uma relação linear entre a variável dependente, Y, e as p variáveis independentes, $X_j,\ j = 1, \dots p$, de certo modo generalizando o modelo (3.3). As condições subjacentes à regressão linear múltipla são análogas às da regressão linear simples. O modelo de regressão linear múltipla seguinte descreve a relação entre as p variáveis independentes, $X_j,\ j = 1, \dots p$. e a variável dependente, Y. Assim consideram-se as seguintes igualdades:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,\ \ i = 1, \dots, n.} \qquad (3.11)$$

onde y_i representa o i-ésimo valor da variável de resposta Y, $x_{i1}, x_{i2}, \dots, x_{ip}$, representam os valores das variáveis preditoras e ε_i representa o erro.

A interpretação dos coeficientes de regressão numa equação de regressão múltipla é mais complexa, pois enquanto que na regressão simples a equação é representada por uma reta, na regressão múltipla esta equação é representada por um plano, se existirem duas variáveis preditoras, ou por um hiperplano, caso existam mais de duas variáveis preditoras.

Neste caso, os coeficientes β_j para $j = 1, \dots p$, representam a variação na variável resposta quando a variável preditora x_j é alterada de uma unidade, permanecendo as outras constantes. (Chatterjee el al, 2012)

3.2.4 Regressão Logística Múltipla

Os modelos logísticos surgiram tanto pela necessidade de modelos mais satisfatórios para o uso de dados qualitativos, como pela dificuldade em aplicar a regressão linear a variáveis dependentes qualitativas. O modelo de regressão logística é o principal modelo utilizado para variáveis dependentes de caracter binário.

A regressão logística é bastante semelhante à regressão linear, em ambos os casos são utilizadas variáveis explicativas para predizer o valor de uma variável resposta, embora na regressão logística a variável resposta tome apenas dois valores possíveis 1 ou 0, ou seja, a presença de uma determinada característica ou a ausência dessa mesma característica. No caso do estudo em questão a variável resposta seja se o cliente converte em apólice $(Y = 1)$ ou, caso contrário, o cliente não converte em apólice $(Y = 0)$.

A principal vantagem da regressão logística face à regressão linear, assenta no facto de esta não impor a existência de linearidade nos parâmetros e por isso, é uma regressão que permite explicar mais aprofundadamente os resultados obtidos, uma vez que a variável resposta é mais precisa.

Uma variável resposta binária, pode ser codificada como tendo dois valores, 0 ou 1, como já referido anteriormente. Mas em vez de prever esses dois valores, iremos tentar modelar as probabilidades de que a variável resposta venha a receber um desses dois valores.

Consideremos um problema de regressão simples em que apenas teremos um preditor, para nos permitir compreender as limitações do modelo de regressão linear. As mesmas considerações serão válidas para o caso da regressão múltipla. Então, seja π a probabilidade de $Y = 1$ quando $X = x$. Se usássemos o modelo linear para descrever π, o nosso modelo para a probabilidade seria dado por:

$$\boldsymbol{\pi = P(Y = 1 | X = x) = \beta_0 + \beta_1 x} \, . \qquad (3.12)$$

Como π é uma probabilidade, terá de estar entre 0 e 1. A função linear dada em (3.12) é ilimitada e, portanto, não poderá ser utilizada para modelar a probabilidade.

A relação entre as probabilidades π e X pode ser representada por uma função de resposta logística. Assemelha-se a uma curva em forma de S, como podemos observar pela imagem seguinte.

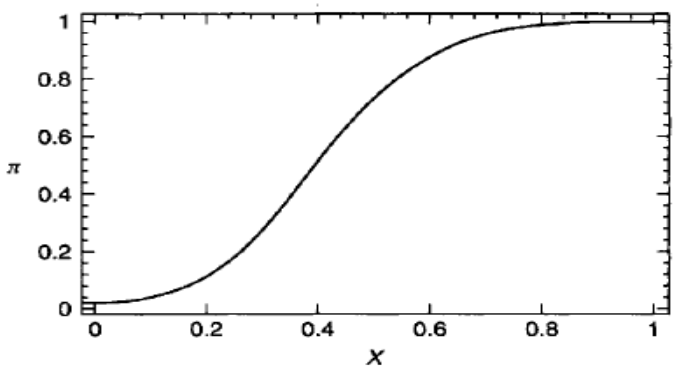

Gráfico 2 - Exemplo da Representação Gráfica do Modelo de Regressão Logística.

Fonte do grífico: Chatterjee el al, 2012

Esta curva poderá ser reproduzida se modelarmos as probabilidades de π e X da seguinte forma:

$$\pi = P(Y = 1 \mid X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \qquad (3.13)$$

onde *e* é a base do logaritmo natural. As probabilidades são modeladas pela função de distribuição da distribuição logística.

O modelo logístico pode ser generalizado diretamente para a situação em que temos várias variáveis preditoras. A probabilidade π é modelada da seguinte forma:

$$\pi = P\left(Y = 1 \mid X_1 = x_1, \dots, X_p = x_p\right) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}} \, . \qquad (3.14)$$

A equação (3.14) é chamada de função de regressão logística. Esta não é linear nos parâmetros $\beta_0, \beta_1, \dots, \beta_p$. No entanto, podem ser linearizados pela transformação *logit*, que em vez de trabalhar diretamente com π, trabalha com um valor transformado de π. Se π é a probabilidade de um evento acontecer, a razão $\frac{\pi}{1-\pi}$ é chamada de *odds ratio* para o evento.

Uma vez que:

$$1-\pi = P\big(Y=0 \mid X_1 = x_1, \dots, X_p = x_p\big) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}, \quad (3.15)$$

o odds-ratio assume a forma $\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$. (3.16)

Aplicando o logaritmo natural a ambos os lados da equação (3.16), obtemos:

$$g\big(x_1, \dots, x_p\big) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \quad (3.17)$$

O logaritmo do *odds ratio* é chamado de *logit*. Pela equação (3.17) podemos observar que a transformação *logit* produz uma transformação linear dos parâmetros $\beta_0, \beta_1, \dots, \beta_p$. (Chatterjee el al, 2012)

3.3 Coeficiente de Determinação – R^2

Embora a inferência estatística para dados bivariados (x_i, y_i) baseada no coeficiente de correlação seja válida apenas para um par binormal, o conceito de correlação tem também aplicabilidade no contexto da regressão tradicional. Uma vez que o coeficiente de correlação mede o grau da relação linear entre as duas variáveis, segue-se que o coeficiente de correlação entre as duas variáveis numa equação de regressão deve estar relacionada com a "qualidade do ajuste" da equação de regressão linear para os pontos de dados amostrais. De facto, o coeficiente de correlação amostral é muitas vezes utilizado como uma estimativa da "qualidade de ajuste" do modelo de regressão. Mais frequentemente, no entanto, o quadrado do coeficiente de correlação entre a variável resposta Y e a preditora X (ou o quadrado do coeficiente de correlação entre a resposta Y e a ajustada $\hat{Y}$) o chamado **coeficiente de determinação**, é usado para este objetivo:

$$R^2 = \frac{\left(\sum(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2} = \frac{\left(\sum(\hat{y}_i-\bar{\hat{y}})(y_i-\bar{y})\right)^2}{\sum(\hat{y}_i-\bar{\hat{y}})^2\sum(y_i-\bar{y})^2}. \quad (3.18)$$

(Freund etal, 2006 e Chatterjee el al, 2012)

Assim, para o caso de regressão linear simples, tem-se que:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \quad (3.19)$$

onde SSR é a soma dos quadrados devido à regressão, e SST é a soma total de quadrados dos desvios relativos à média numa análise de regressão linear simples:

$$SST = \sum(y_i - \bar{y})^2 \quad e \quad SSR = \sum(\hat{y}_i - \bar{y})^2. \quad (3.20 \text{ e } 3.21)$$

Note-se que denotando a soma dos quadrados dos resíduos ou erros por:

$$SSE = \sum(y_i - \hat{y}_i)^2, \quad (3.22)$$

tem-se que a usual partição de quadrados:

$$SST = SSR + SSE. \tag{3.23}$$

Tem-se que $0 \leq R^2 \leq 1$, como decorre de imediato da desigualdade $SSR \leq SST$. Se R^2 está perto de 1, então X é responsável por uma grande parte da variação em Y.

Assim, o coeficiente de determinação R^2 é uma medida descritiva da relativa robustez correspondendo à regressão, e representa a redução proporcional da variação total de associada à regressão e é amplamente utilizado para descrever a eficácia de um modelo de regressão linear. O coeficiente R^2 tem uma interpretação similar para o caso de regressão múltipla.

Em regressão linear múltipla, tem-se igualmente que o grau de relação linear entre a variável dependente Y e o conjunto de preditoras $X_1, X_2, \dots, X_p$ pode ser avaliado através da análise do diagrama de dispersão de y_i versus $\hat{y}_i$ e o coeficiente de correlação dado por:

$$Cor(Y, \hat{Y}) = \frac{\sum(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum(\hat{y}_i - \bar{\hat{y}})^2 \sum(y_i - \bar{y})^2}}, \tag{3.24}$$

onde $\bar{y}$ e $\bar{\hat{y}}$ denotam, tal como anteriormente, respetivamente a média dos valores $y_i$da resposta Y e a média dos valores ajustados $\hat{y}_i$. Tal como no caso de regressão simples, o *coeficiente de determinação* é dado por :

$$R^2 = \left[Cor(Y, \hat{Y})\right]^2 \tag{3.25}$$

que pode ser expresso como:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \tag{3.26}$$

com notações paralelas às anteriores para regressão simples.

Assim, R^2pode ser interpretado como a proporção da variabilidade total na variável de resposta Y que pode ser explicada pelo conjunto de variáveis preditoras $X_1, X_2, \dots, X_p$. (Chatterjee el al, 2012)

3.4 Teste de *Wald*

Na análise da conversão dos clientes iremos ainda recorrer a este tipo de testes.

O teste de Wald resulta da comparação entre a estimativa de máxima verosimilhança do parâmetro ($\widehat{\beta\imath}$) e a estimativa do seu erro padrão.

As hipóteses a serem testadas são: $H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0$ com $j = 0, \dots, p$ sendo o p o número de variáveis em estudo.

A estatística de teste para a seleção da variável β_j na regressão logística, sob H_0 é dada por:

$$W_j = \frac{\widehat{\beta}_i}{\widehat{DP(\beta_i)}}. \qquad (3.28)$$

O valor do *p-value* é calculado da seguinte forma: $P(|Z| > |w_j|)$, onde w_j é o valor observado da estatística de teste e Z a variável aleatória com distribuição normal padrão.

3.5 Matriz de Confusão

A matriz de confusão é uma tabela que nos permite obter uma visualização simplificada do número de classificações corretas e do número de classificações preditas de uma classe (Han e Kamber, 2006).

Uma forma de avaliação da qualidade de ajustamento do modelo, passa pela comparação dos valores previstos com os valores observados. Quanto maior for o nível de concordância entre os valores observados e previstos, melhor será o modelo.

Chama-se valores positivos aos valores iguais à unidade e valores negativos àqueles que são nulos. Os valores positivos que são previstos como tal chamam-se verdadeiros positivos (VP) e os que são previstos de forma errada como negativos são denominados como falsos negativos (FN). Da mesma forma, os verdadeiros valores negativos podem ser previstos corretamente como sendo negativos, isto é, os verdadeiros negativos (VN), ou podem ser previstos como positivos, denominando-se falsos positivos (FP). Após a classificação de todos os pares de valores, tanto observados como previstos, os resultados da contagem do total de pares nas classes VP, FN, VN e FP são apresentados através da Matriz de Confusão.

	Previstos		
Observados	VERDADEIRO	FALSO	TOTAL
VERDADEIRO	VP	FN	VP+FN
FALSO	FP	VN	FP+VN
TOTAL	VP+FP	FN+VN	nº observações

Tabela 2 – Matriz de Confusão

O desejado é que os números totais de falsos positivos (FP) e de falsos negativos (FN) sejam os menores possíveis. Estes valores podem ser alertados, se o ponto de corte for modificado. Mas em contrapartida, com a modificação do ponto de corte não é possível diminuir o número e falsos positivos sem aumentar os falsos negativos, e vice-versa (Alpuim, 2017).

Com o objetivo de verificar a qualidade do modelo, podemos definir algumas probabilidades:

- **Sensibilidade**: probabilidade de uma observação ser classificada como positiva dado que é positiva.

$$Sensibilidade = \frac{VP}{VP+FN} \quad (3.29)$$

- **Especificidade**: probabilidade de uma observação ser classificada como negativa dado que é negativa.

$$Especificidade = \frac{VN}{VN+FP} \quad (3.30)$$

A **precisão** é a proximidade entre o resultado de uma medida e o seu valor verdadeiro, ou seja, dá-nos uma aproximação da probabilidade de previsões corretas, sejam elas positivas ou negativas:

$$\boldsymbol{Precisão} = \frac{VP+VN}{nº\ de\ observações} \quad (3.31)$$

3.6 Métodos de Seleção de Variáveis

Tradicionalmente o objetivo da construção de modelos estatísticos passa por encontrar o modelo mais simples que melhor explica os dados. Existem algumas técnicas para a seleção de variáveis.

Foram analisados três métodos de seleção de variáveis.

3.6.1 Método de Seleção *Stepwise*

O método de seleção de variáveis ***Stpewise*** inicia-se com zero variáveis e verifica em cada inclusão de uma nova variável a importância das variáveis já presentes no modelo sendo possível a remoção de alguma variável. O algoritmo termina quando não existem mais variáveis para entrar ou sair do modelo.

Qualquer procedimento *Stpewise* para seleção ou exclusão de variáveis, é baseado num algoritmo estatístico que verifica a 'importância' das variáveis e as inclui ou exclui do modelo, com base numa regra de decisão fixa. A 'importância' de cada variável é definida pela significância estatística do seu coeficiente. A estatística usada depende das suposições do modelo. Na regressão linear, é utilizado o teste *F*, uma vez que os erros são considerados normalmente distribuídos. Por outro lado, na regressão logística, os erros seguem uma distribuição binomial e por isso, a significância das variáveis é avaliada por um teste de Razão de Verosimilhanças em cada passo, como passaremos a descrever.

O método de seleção de variáveis *Stepwise* é descrito pelos seguintes passos:

Passo (0): Considere-se que temos *p* variáveis independentes, todas elas consideradas importantes para explicar a variável resposta. O passo (0) inicia-se com um ajuste ao modelo que contém apenas os termos independentes e é calculado o valor do logaritmo de verosimilhança, L_0.

Em seguida, para cada uma das *p* variáveis independentes ajustam-se vários modelos univariados e comparam-se os respectivos logaritmos de verosimilhança, L_0. O valor do logaritmo de verosimilhança do modelo que contém a variável x_j, no passo (0) é designado por $L_j^{(0)}$. O subscrito j refere-se à variável que foi adicionada ao modelo e o sobrescrito 0 refere-se ao passo em que nos encontramos. Esta notação será utilizada ao longo de todo o processo do método *Stepwise* para acompanhar tanto o número de passos como também o número de variáveis presentes no modelo.

Para o modelo que contém a variável x_j *versus* o modelo que contém somente os termos constantes, temos que o valor do teste da razão de verosimilhanças é dado por: $G_j^{(0)} = -2\left(L_0 - L_j^{(0)}\right)$, e o seu *p-value* é designado por $p_j^{(0)}$. Portanto, o valor do *p-value* é dado pela probabilidade $P\left[\chi^2(v) > G_j^{(0)}\right] = p_j^{(0)}$, onde $v = 1$ se x_j é uma variável contínua e $v = k - 1$ se x_j é uma variável policotómica com k categorias.

A variável mais importante é aquela que apresenta um *p-value* mais reduzido. Se designarmos esta variável por x_{e_1}, temos $p_{e_1}^{(0)} = min\left(p_j^{(0)}\right)$. Sendo que o p_e indica o nível de entrada das variáveis no modelo. O subscrito e_1 é utilizado para indicar que a variável é candidata a entrar no passo (1).

Passo (1): Este passo inicia-se com um ajuste ao modelo de regressão logística que contém a variável x_{e_1}. Seja $L_{e_1}^{(1)}$ o logaritmo da verosimilhança para este modelo. Uma vez que o modelo já contém a variável x_{e_1}, para determinar se qualquer umas das restantes $p-1$ variáveis é importante, ajustamos os modelos de regressão logística $p-1$ contendo a variável x_{e_1} e $x_j, j = 1, \dots, p$ e $j \neq e_1$. Designamos $L_{e_1 j}^{(1)}$, por o logaritmo da verosimilhança que contém x_{e_1} e x_j. Temos então que a estatística do qui-quadrado do modelo é dada pela seguinte expressão: $G_j^{(1)} = -2\left(L_{e_1}^{(1)} - L_{e_1 j}^{(1)}\right)$. O valor do *p-value* para esta estatística é dado por $p_j^{(1)}$. Seja x_{e_2} a variável que

possuía o menor *p-value* neste passo, onde $p_{e_2}^{(1)} = min(p_1^j)$. Então, se $p_j^{(1)} < p_e$ prosseguiremos para o passo (2), caso contrário paramos neste passo.

Passo (2): Este passo começa com um ajuste ao modelo que contém as variáveis x_{e_1} e x_{e_2}. É possível que, uma vez adicionada a variável x_{e_2}, a variável x_{e_1}possa ter perdido a sua importância para o modelo. Assim, este passo inclui uma verificação de variáveis regressiva, onde é excluída alguma variável se necessário. Seja $L_{-e_j}^{(2)}$ o valor do logaritmo de verosimilhança do modelo após ser removida a variável x_{e_1}. Do mesmo modo, a estatística do qui-quadrado do modelo *versus* o modelo completo no passo (2) é dada por: $G_{-e_j}^{(2)} = -2\left(L_{-e_j}^{(2)} - L_{e_1e_2}^{(2)}\right)$ e $p_{-e_j}^{(2)}$ corresponde ao valor do *p-value*. Admitindo que x_{r_2} foi a variável excluída do modelo, pois possuía um *p-value* bastante elevado, *p-value* esse definido por: $p_{r_2}^{(2)} = max\left(p_{-e_j}^{(2)}, p_{-e_2}^{(2)}\right)$. Para decidir se a variável x_{r_2} deverá ser ou não removida do modelo, irar-se-á comparar o valor de $p_{r_2}^{(2)}$ com $p_{r_2}^{(2)}$sendo que p_R indica o nível de remoção de variáveis do modelo. Seja qual for o valor escolhido para p_R este deverá ser sempre maior que o valor de p_E para evitar a remoção da mesma variável em etapas sucessivas. Então, se $p_{r_2}^{(2)} < p_R$, a variável x_{r_2} será removida do modelo, caso contrário, irá permanecer no modelo.

Na fase de seleção progressiva, serão ajustados $p - 2$ modelos, contendo as variáveis x_{e_1}, x_{e_2}e x_j, onde $j = 1, \dots, p, j \neq e_1, e_2$. Avalia-se o logaritmo de verosimilhança para cada um dos $p - 2$ modelos e determina-se a estatística de razão de verosimilhança entre os novos modelos e o modelo que apenas contém as variáveis x_{e_1} e x_{e_2}, calculando os respectivos *p-value*. Seja x_{e_3} a variável que apresenta o menor *p-value*, ou seja, $p_{e_3}^{(2)} = min\left(p_j^{(2)}\right)$. Então, se $p_{e_3}^{(2)} < p_e$ prossegue-se para o passo (3), caso contrário, o processo termina.

Passo (3): Este terceiro passo é bastante idêntico ao passo anterior. O processo irá continuar até ao último passo, o passo (S).

(…)

Passo (S): Neste último passo, todas as variáveis presentes no modelo apresentam um *p-value* inferior a p_R e todas as restantes variáveis que não foram incluídas no modelo possuem um p-value superior a p_e. Então, o modelo nesta fase contém todas as variáveis que foram consideradas importantes segundo os valores de p_R e p_e. (Hosmer & Lemeshow, 2000)

3.6.2 Método de Seleção *Forward*

O método de seleção de variáveis ***Forward***, também conhecido como método da seleção progressiva, é um método de inclusão de variáveis. O *Forward* parte do princípio que não existe nenhuma variável no modelo e em cada iteração irá testar a inclusão de cada variável juntamente com as variáveis previamente selecionadas, a fim de identificar a variável com maior significância de entre as restantes. Em seguida deve ser testada a significância dessa variável de forma a decidir se deve ou não ser incluída no modelo. Este processo deve ser repetido até que exista alguma variável que deixe de ser significativa (Alpuim, 2017).

3.6.3 Método de Seleção *Backward*

O método de seleção de variáveis ***Backward***, também conhecido como método de seleção regressiva, começa por incluir todas as variáveis no modelo e depois, elemina as variáveis progressivamente. Depois da variável menos significativa ser eliminada, o modelo é reajustado com as restantes variáveis, repetindo-se o processo até que apenas fiquem as variáveis significativas (Alpuim,2017).

3.7 Critério de Informação de *Akaike*

Os métodos de seleção podem basear-se em diversos critérios sendo o Critério de Informação de *Akaike* (AIC) um dos mais utilizados. Este critério surgiu em 1974 por Hirotugu Akaike.

O Critério de Informação de *Akaike* é um critério que avalia a qualidade do ajuste do modelo estimado pelo método da máxima verosimilhança.

Dado um conjunto de dados e vários modelos concorrentes, pode-se classificá-los de acordo com o seu AIC, considerando como os melhores, aqueles que possuem valores mais reduzidos de AIC.

O critério de Informação de *Akaike* é definido como:

$$\boldsymbol{AIC} = -2\,\mathbf{log}(\boldsymbol{L}) + 2\boldsymbol{p} \qquad (3.31)$$

onde, L corresponde à função de máxima verosimilhança do modelo e p ao número de variáveis explicativas consideradas no modelo.

3.8 Curva de ROC e *Area Under Curve* (AUC)

A curva de ROC (*Receiver Operating Characteristic*) é uma técnica gráfica, a curva é obtida através do cálculo da sensibilidade e da especificidade para cada ponto de corte, representando-se graficamente os pontos de coordenadas (1-especificidade, sensibilidade). A sensibilidade é representada no eixo das ordenadas enquanto que 1-espeficidade é representada no eixo das abcissas, variando ambas entre 0 e 1. Um modelo é considerado perfeito quando a sensibilidade e a especificidade são iguais a 1.

A curva de ROC pode ser utilizada na comparação de modelos, uma vez que a exatidão de um teste diagnóstico é proporcional à área abaixo da curva, logo quanto maior for a área abaixo da curva maior será a exatidão do modelo.

A área abaixo da curva de ROC (AUC – *Area Under the (ROC) Curve*) é um dos índices de precisão mais utilizados para avaliar a qualidade da curva e o desempenho do teste de diagnóstico.

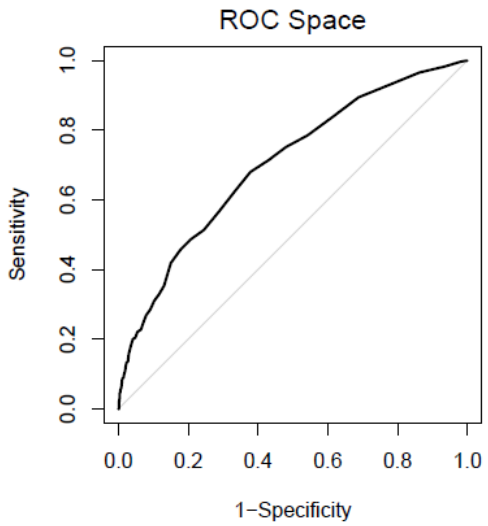

Gráfico 3 - Exemplo Curva de *Roc*

Fonte do gráfico: Tese Guilherme Mousinho

Quanto maior o valor do AUC maior será a exatidão do modelo. Os valores do AUC podem ser avaliados em 5 níveis discriminatórios:

Valor	Classificação
[0.9 ; 1]	**Discriminação Excecional**
[0.8 ; 0.9]	**Discriminação Excelente**
[0.7 ; 0.8]	**Discriminação Aceitável**
[0.6 ; 0.7]	**Discriminação Fraca**
[0.5 ; 0.6]	**Não existe Discriminação**

Tabela 3 - Classificação AUC

3.9 Teste do Qui-Quadrado

Através de uma tabela de contingência é possível calcular uma estatística a partir da qual podemos efetuar um teste de hipóteses designado de teste de Qui-Quadrado, a fim de averiguar se as variáveis são independentes.

As hipóteses a serem testadas são as seguintes:

$$H_0 : \textit{as variáveis X e Y são independentes}$$

$$vs.$$

$$H_1 : \textit{as varáveis X e Y não são independentes}$$

O valor da estatística de teste é dado pela seguinte expressão:

$$X^2 = \sum_{i,j=1}^{lc} \frac{\left(O_{ij} - \hat{e}_{ij}\right)^2}{\hat{e}_{ij}}, \qquad (3.32)$$

onde $\hat{e}_{ij}$ representa a estimativa da frequência esperada sob a hipótese nula de independência, e O_{ij} a frequência observada para a célula ij sendo l e c o número de categorias em que se classificaram as variáveis X e Y, respetivamente.

Quando o número de observações n é elevado a distribuição da estatística X^2 é aproximadamente a do χ^2 com $v = (l-1)(c-1)$ graus de liberdade (g.l.); sendo assim, para um valor x^2 obtido da estatística de Qui-Quadrado, um valor aproximado do *p-value* é dado por $P[\chi_v^2 > x^2]$.

3.10 Coeficiente de Correlação de *V-Cramer*

O coeficiente de *V-Cramer* é uma medida de associação entre duas variáveis medidas numa escala categórica, ou seja, este pode ser aplicado em situações onde a informação se encontra distribuída por categorias nominais não ordenáveis. Este coeficiente é obtido diretamente a partir da estatística χ^2.

$$V = \sqrt{\frac{\chi^2/n}{(k-1)}} = \sqrt{\frac{\chi^2}{n(k-1)}}, \qquad (3.33)$$

onde *n* representa o número total de observações e *k* representa o mínimo entre o número de linhas e colunas da tabela de contingência, isto é, $k = \min(l, c)$.

O coeficiente de *V-Cramer* pode tomar valores entre 0 e 1, onde o valor 0, corresponde á ausência de associação entre as variáveis, valores próximos de zero correspondem a fraca associação entre as variáveis, e contrariamente, valores próximos de 1 correspondem a uma elevada associação entre as variáveis.

Capítulo 4: Modelo de Conversão

Para a construção do modelo foram consideradas todas as simulações relativas à nova tarifa automóvel da companhia, implementada em setembro de 2017, sendo assim o período temporal da base de dados em estudo será de 26 de setembro de 2017 a 14 de janeiro de 2019.

Apenas serão contempladas simulações de clientes individuais, em que o objeto seguro será um veículo ligeiro, um veículo misto ou uma caminheta. O modelo não irá contemplar simulações de frotas, protocolos ou ordens profissionais.

4.1 Preparação dos Dados

Uma das fases mais importantes quando se trabalha com uma elevada quantidade de dados, é a preparação da base de dados. Por preparação da base de dados entendemos a 'limpeza' dos dados originais. Este processo exige bom senso e algum conhecimento do estudo em causa, por forma a permitir uma correta limpeza dos dados. Este processo inclui procedimentos como a remoção de dados inconsistentes e o tratamento de valores omissos.

Um modelo estatístico é tão bom quanto os dados subjacentes. Consequentemente, uma boa compreensão dos dados é um ponto de partida essencial para a modelação.

Para o problema em causa, uma vez que existiam muitas simulações por chave, chave esta composta por o número de contribuinte do cliente, a matrícula do carro seguro e o código de identificação do agente, houve a necessidade de fazer uma maior compactação dos dados.

Para uma chave que possuísse mais que uma simulação num intervalo de 60 dias, apenas contaria uma simulação, a mais atual ou a que obtivesse conversão. A escolha deste intervalo de 60 dias foi feita tendo em consideração que uma simulação na companhia é válida por 30 dias.

Como podemos ver pelo exemplo seguinte, um cliente que apresente cinco simulações com a mesma chave, para *x* meses, por cada intervalo de 60 dias apenas contará uma simulação, no caso do primeiro intervalo onde existem 3 simulações no período de 60 dias apenas contará uma, ou a que obteve conversão, ou caso não se tenha obtido conversão, a simulação mais atual.

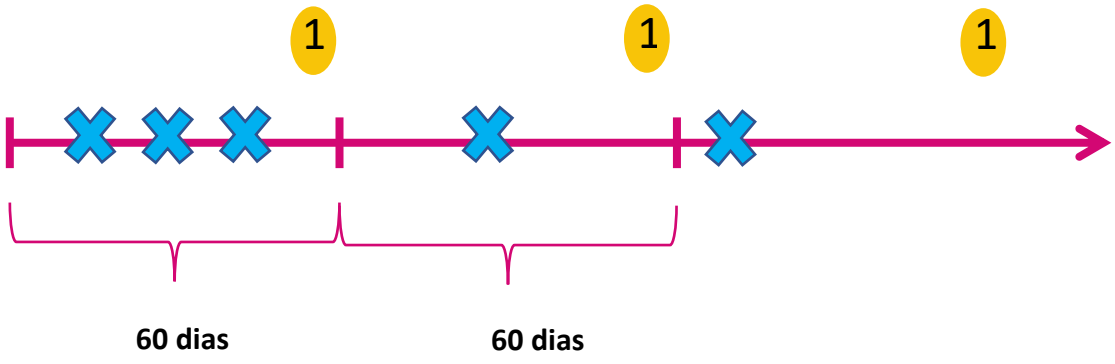

Figura 2 – Exemplo Compactação dos Dados

4.2 Variáveis em Estudo

Através das simulações realizadas pelos clientes, são recolhidos pela seguradora, inúmeros dados acerca do cliente e do seu objeto seguro, dando assim origem à construção das variáveis que serão utilizadas no modelo. Estas variáveis encontram-se agrupadas em 6 categorias:

- Informação Geográfica
- Informação do Cliente
- Informação do Agente
- Características do Veículo
- Características da Apólice
- Variação do Prémio

Temos como possíveis variáveis explicativas, por categoria:

- **Informação Geográfica**
    - Urbano-Rural – Classificação da morada do tomador do seguro, como urbana ou rural de acordo com os dados CAMEO;
    - Concelho – Concelho da morada do tomador do seguro;
    - Distrito – Distrito da morada do tomador do seguro;
    - Zona_RC – Indica a zona de risco de responsabilidade civil, com uma escala de 1 a 30, sendo o 1 o mais gravoso;
    - Zona_DP - Indica a zona de risco de danos próprios, com uma escala de 1 a 30, sendo o 1 o mais gravoso.

- **Informação do Cliente**
    - Idade do condutor – Idade do condutor do veículo seguro;
    - Idade da carta de condução – Idade da carta de condução do condutor do veículo seguro;
    - Tipo de Cliente – Indica o tipo de cliente: Particular ou Empresa;
    - Sexo – Indica o género do cliente, neste caso existem três géneros: homem, mulher ou empresa; esta variável não será utilizada uma que não se pode fazer distinções de género;
    - Jovem – Indica se o cliente é jovem, ou seja, se o cliente possui menos de 25 anos;

- Cliente Novo – Indica se o cliente é novo para a companhia, ou se caso contrário, já possuiu alguma apólice automóvel na companhia, num período de 5 anos;

- **Informação do Agente**
  - Rede e Canal Comercial – Rede e Canal Comercial correspondentes aos agentes;
  - Código do Agente – Código de identificação do agente;

- **Características do Veículo**
  - Categoria – Categoria do veículo seguro;
  - Matrícula – Matrícula do veículo seguro;
  - Idade do Veículo – Idade do veículo seguro;
  - Cilindrada – Cilindrada do veículo seguro;
  - Tara – Tara do veículo seguro;
  - Potência – Potência do veículo seguro;
  - Peso Bruto – Peso bruto do veículo seguro;
  - Marca – Marca do veículo seguro;
  - Combustível – Tipo de combustível do veículo seguro;
  - Caixa de Velocidades – Caixa de velocidades do veículo seguro: manual ou automática;
  - Número de Portas – Número de portas do veículo seguro;
  - Peso Potência – Corresponde à divisão do valor da tara pelo valor da cilindrada do veículo seguro;
  - Controlo de Travagem – Indica se o veículo seguro possuiu este equipamento de qualidade, em caso afirmativo recebe 5% de desconto no valor do prémio;
  - Melhoria de Visibilidade - Indica se o veículo seguro possuiu este equipamento de qualidade, em caso afirmativo recebe 5% de desconto no valor do prémio;
  - Controlo de Condução - Indica se o veículo seguro possuiu este equipamento de qualidade, em caso afirmativo recebe 5% de desconto no valor do prémio;
  - Alarme de Segurança - Indica se o veículo seguro possuiu este equipamento de qualidade, em caso afirmativo recebe 5% de desconto no valor do prémio.

- **Características da Apólice**
    - Data da Simulação – Data em que foi realizada a simulação;
    - Bónus *Malus* – Indica o escalão de bónus *malus* em que o cliente se encontra;
    - Pack – Pack escolhido pelo cliente;
    - **Conversão** – Indica se o cliente converteu em apólice ou não; (**variável dependente**);
    - Cobrança Bancária – Forma de cobrança, que poderá ser feita por transferência bancária ou não.

- **Variação do Prémio**
    - Desconto Comercial – Indica se o cliente tem ou não desconto comercial;
    - Prémio Comercial – Prémio com desconto e com bónus *malus*.

Poderiam ter sido utilizadas mais variáveis, no entanto existem dados que não são fiáveis, ou por um elevado número de *missing values* ou por um mau preenchimento.

4.3 Análise Descritiva das Variáveis

De forma a não violar o acordo de confidencialidade para com a companhia e de forma a não comprometer os dados da empresa, não será possível apresentar os valores das taxas de conversão nem as suas respetivas exposições. No entanto, será possível analisar as tendências e daí tirar as devidas conclusões.

- **Informação Geográfica**
    - **Urbano/Rural**

A variável Urbano/Rural representada no gráfico seguinte é uma variável externa à companhia, que nos permite compreender através dos códigos postais dos clientes, se estes se encontram numa zona urbana ou rural. Como podemos observar, é na zona rural que se observa um maior número de clientes tanto a simular como a converter.

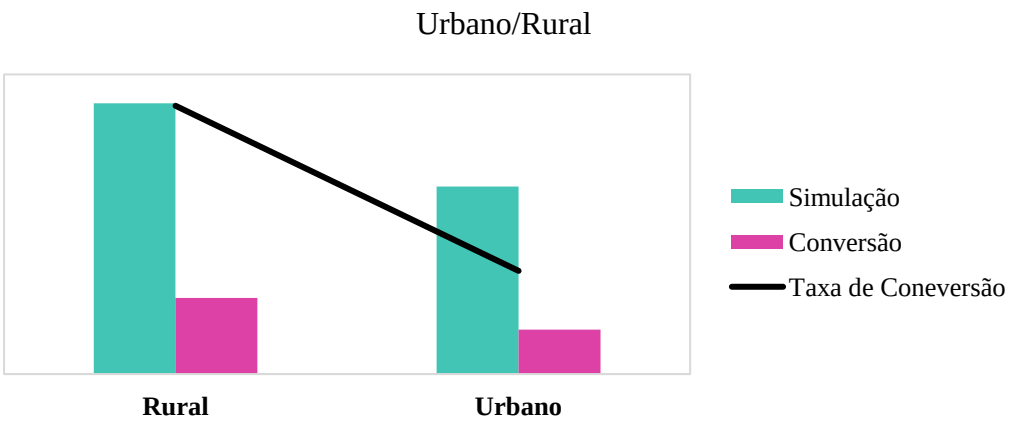

Gráfico 4 – Taxa de Conversão da Variável Urbano/Rural

- **Distrito**

No gráfico 6, podemos observar a variação da taxa de conversão automóvel por distrito, podemos reparar que os distritos que apresentam uma cor mais escura, ou seja, uma taxa de conversão mais elevada são Beja e Portalegre. No entanto para complementar, o gráfico 5 demonstra que apesar da taxa de conversão ser mais elevada nos dois distritos anteriormente descritos, é no Porto, Lisboa e Aveiro que se encontra a maior parte da carteira automóvel.

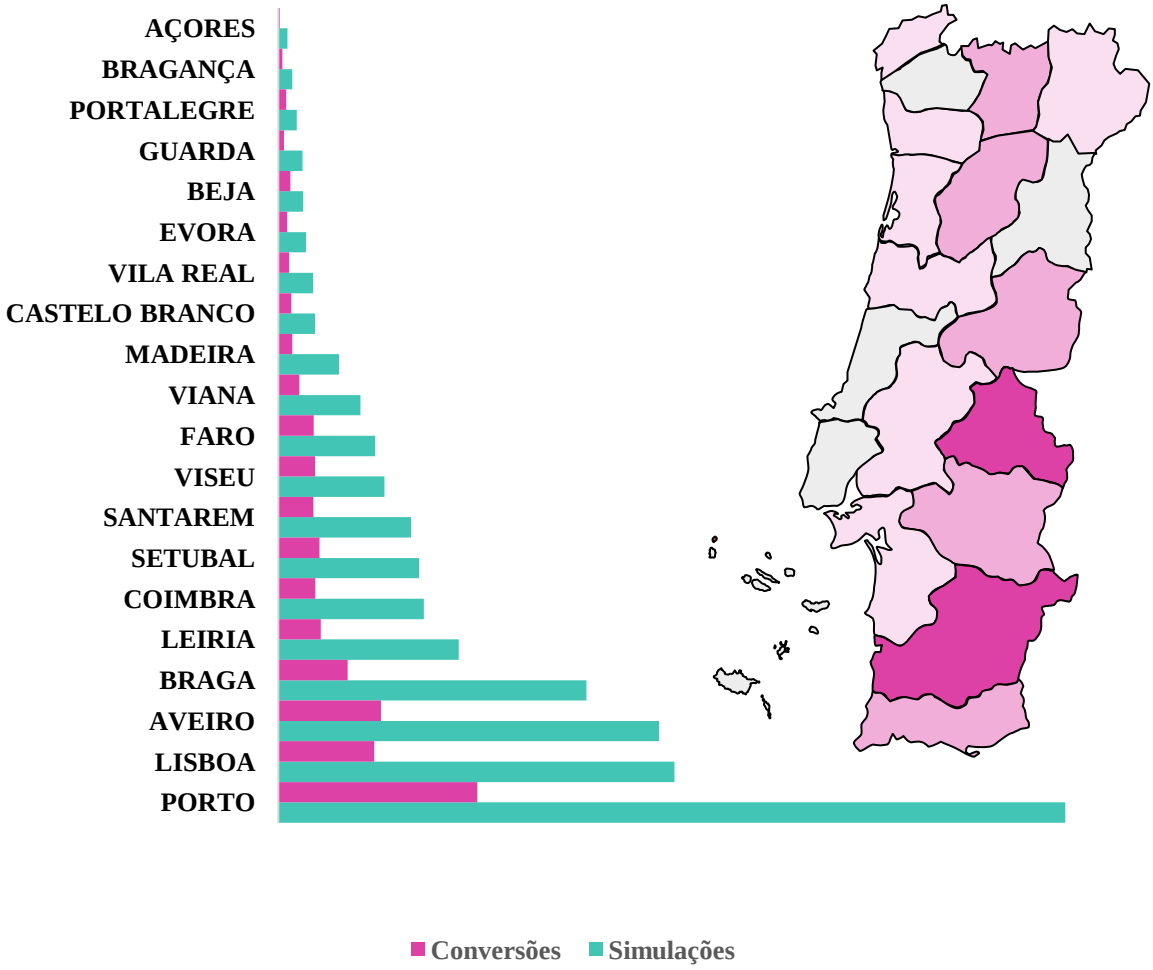

Gráfico 5 – Simulações versus Conversões da Variável Distrito

Gráfico 6 – Taxa de Conversão da Variável Distrito

> 40%	25% - 30%
30% - 40%	< 25%

Figura 3 - Legenda do Gráfico 5

- **Zona RC**

Através do seguinte gráfico podemos constatar que nas zonas mais gravosas, ou seja, nas zonas onde os valores são mais reduzidos, podemos observar uma taxa de conversão mais ou menos constante, no entanto é nas zonas menos gravosas que se registam as taxas de conversão mais elevadas como já seria de esperar. Sendo ainda salientar um pico na zona 26, que apresenta a taxa de conversão mais reduzida.

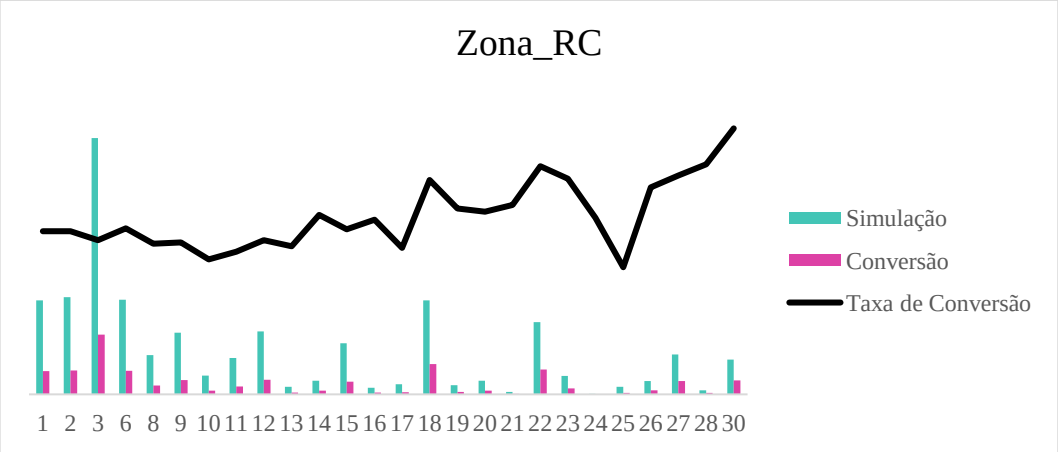

Gráfico 7 - Taxa de Conversão da Variável Zona_RC

- **Zona DP**

O gráfico 8 mostra-nos a variação da taxa de conversão automóvel por zona DP, como seria de esperar as taxas de conversão são mais elevadas em zonas menos gravosas.

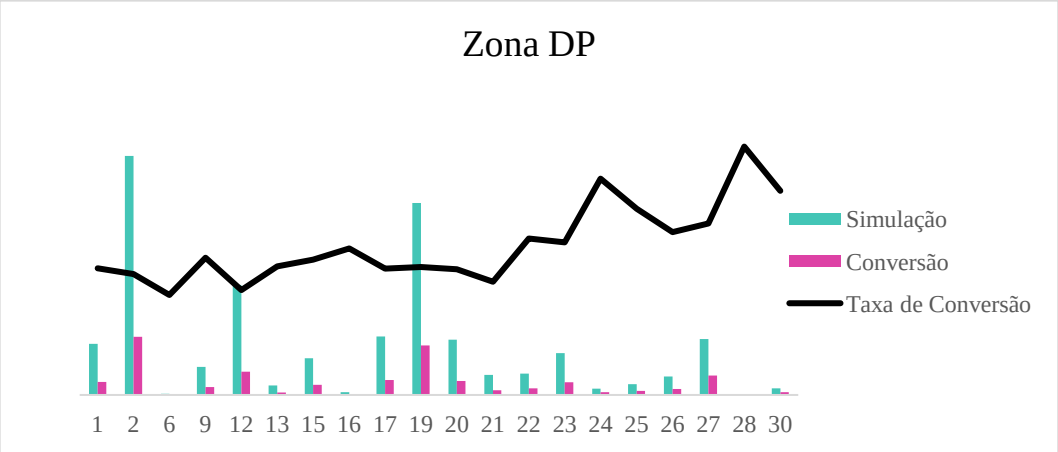

Gráfico 8 - Taxa de Conversão da Variável Zona_DP

- **Informação do cliente**
    - Idade do condutor

O gráfico 9 mostra-nos a variação da taxa de conversão automóvel pela idade da pessoa segura. Verificamos que os segurados que mais simulam e por consequência, os que mais convertem encontram-se entre os 40 e os 50 anos de idade. A partir dos 73 anos podemos observar uma baixa expressão nos dados relativos às simulações, pelo que a partir dessa idade que a taxa de conversão se apresenta mais elevada. Condutores até aos 25 anos, são considerados condutores 'jovens' e deste modo, condutores de risco.

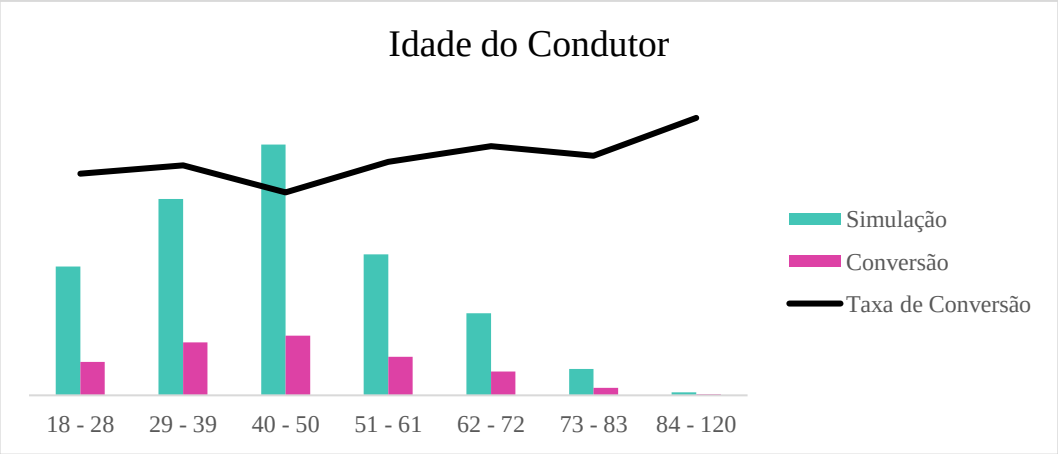

Gráfico 9 - Taxa de Conversão da Variável Idade do Condutor

- Idade da carta de condução

O gráfico 10 apresenta-nos a variação da taxa de conversão automóvel pela idade da carta da pessoa segura, quanto mais elevada for esta variável maior será a experiência do condutor. Podemos verificar que é entre os 21 e os 25 anos de carta que estão representados o maior número de registos.

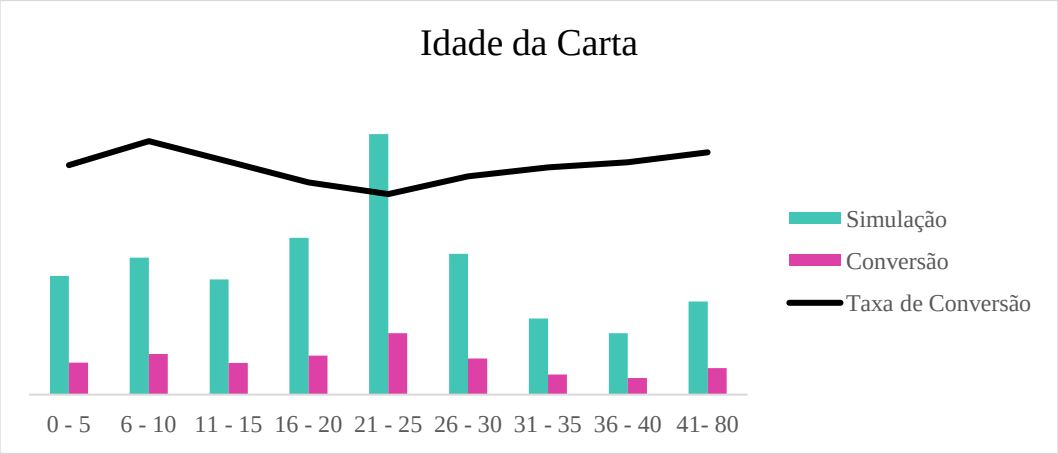

Gráfico 10 - Taxa de Conversão da Variável Idade da Carta de Condução

- Tipo de Cliente

Este gráfico representa a variação da taxa de conversão automóvel pelo tipo de cliente automóvel, é nos clientes particulares que se encontra a maior parte da carteira de negócio.

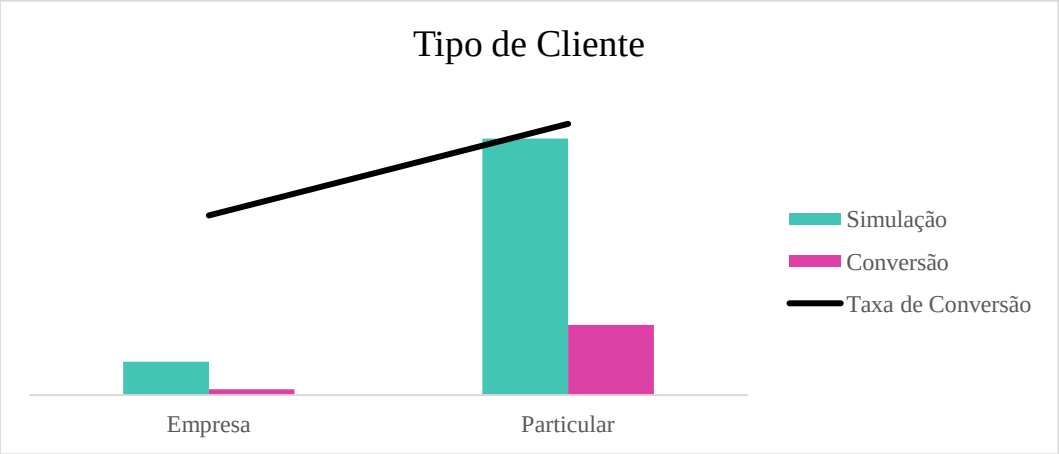

Gráfico 11 - Taxa de Conversão da Variável Tipo de Cliente

- Jovem

É nos jovens que está concentrada a maior preocupação das seguradoras, uma vez que estes são considerados condutores de risco por possuírem pouca experiência e por isso taxas de sinistralidade mais elevadas. Uma vez que o prémio destes também sofre um agravamento é de esperar que não seja nos jovens que está concentrada a taxa mais elevada de conversão automóvel.

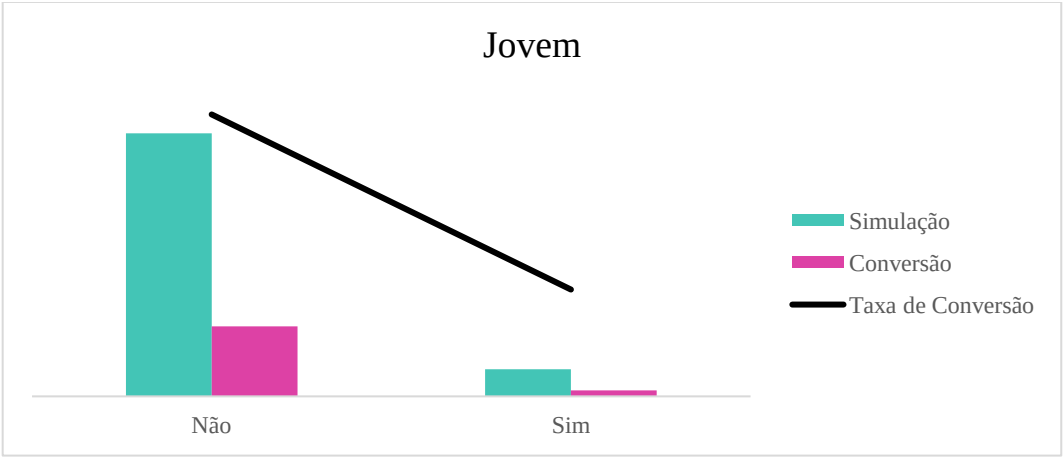

Gráfico 12 - Taxa de Conversão da Variável Jovem

- Cliente Novo

Pelo gráfico seguinte podemos observar a variação da taxa de conversão automóvel pela variável de cliente novo, ou seja, estes são considerados clientes que num período de 5 anos não possuíram uma apólice na companhia. São os clientes novos os que mais simulam, no entanto, são estes que apresentam uma taxa de conversão mais reduzida.

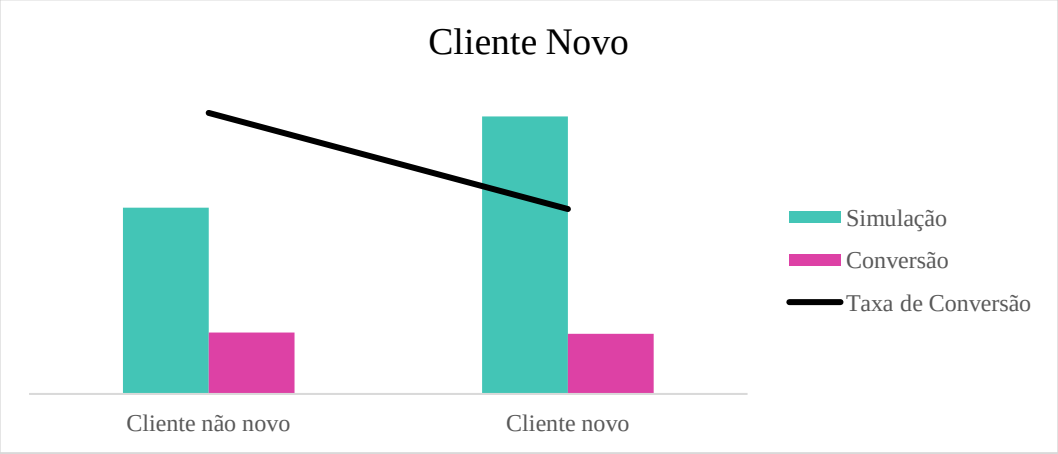

Gráfico 13 - Taxa de Conversão da Variável Cliente Novo

- **Informação do agente**
  - Canal e Rede Comercial

Através do seguinte gráfico podemos observar a variação da taxa de conversão automóvel por o Canal e Rede Comercial do agente, podemos concluir que são os agentes multimarca, ou seja, os que vendem apólices de diversas seguradoras os que mais convertem, no entanto é nos exclusivos, aqueles que trabalham apenas com a marca AGEAS os que possuem taxas de conversão automóvel mais elevadas.

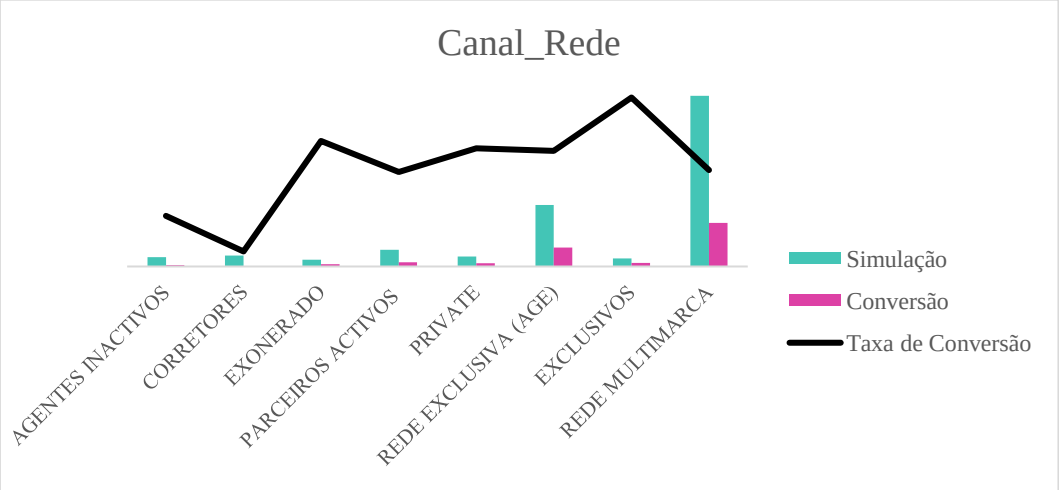

Gráfico 14 - Taxa de Conversão da Variável Canal_Rede

- **Características do veículo**
  - Categoria

No gráfico seguinte podemos analisar a variação da taxa de conversão automóvel de acordo com a categoria do veículo seguro. Podemos verificar que a maior parte das apólices em carteira, ou seja, a maioria das conversões, dizem respeito a um veículo ligeiro, mas no entanto, são os veículos mistos que apresentam uma taxa de conversão mais elevada.

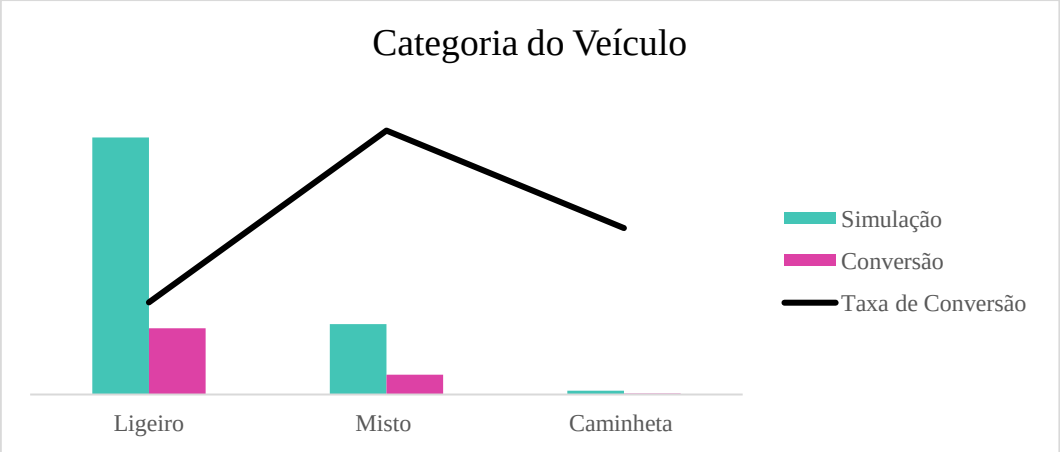

Gráfico 15 - Taxa de Conversão da Variável Categoria do Veículo

  - Idade do Veículo

No gráfico seguinte podemos observar a variação da taxa de conversão automóvel pela idade do veículo seguro. É possível verificar que a taxa de conversão se encontra mais elevada a partir dos 21 anos de idade do veículo e contínua aproximadamente constante para os anos seguintes.

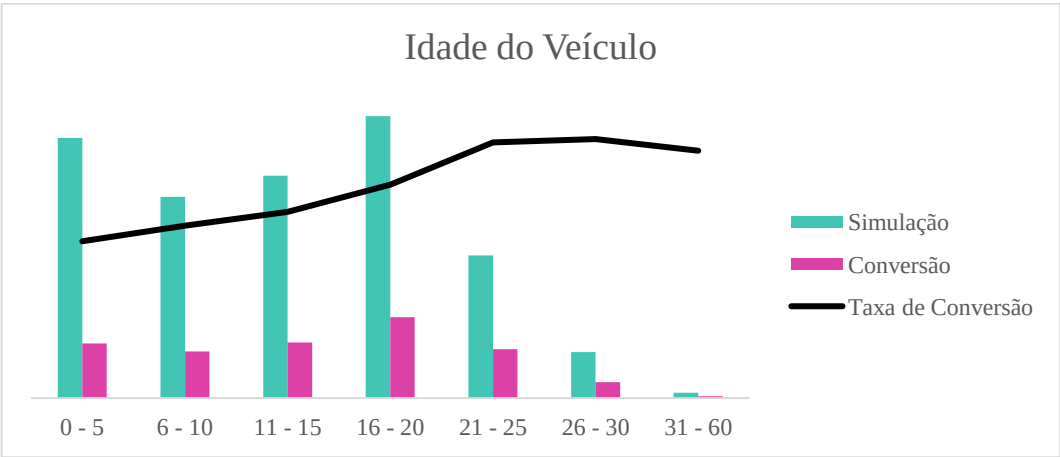

Gráfico 16 - Taxa de Conversão da Variável Idade do Veículo

- Cilindrada

O gráfico 17 mostra-nos a variação da taxa de conversão automóvel pela cilindrada do veículo seguro. A maior parte da carteira encontra-se entre os 1001 e os 2000. Podemos concluir que é nos valores de cilindrada mais reduzidos que a taxa de conversão é mais elevada.

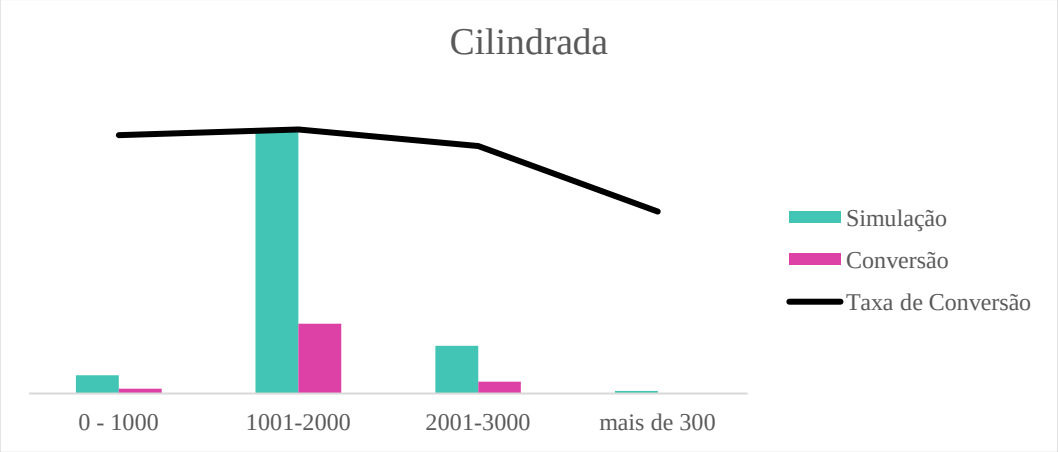

Gráfico 17 - Taxa de Conversão da Variável Cilindrada

- Tara do veículo

No gráfico seguinte podemos observar a variação da taxa de conversão automóvel pela tara do veículo seguro. A tara corresponde ao peso do veículo, sem passageiros nem carga. Pelo que, podemos observar que a maior parte da certeira se encontra entre os 1001 e os 1500, mais uma vez podemos constatar que a taxa de conversão é mais elevada quando os valores da tara são mais reduzidos.

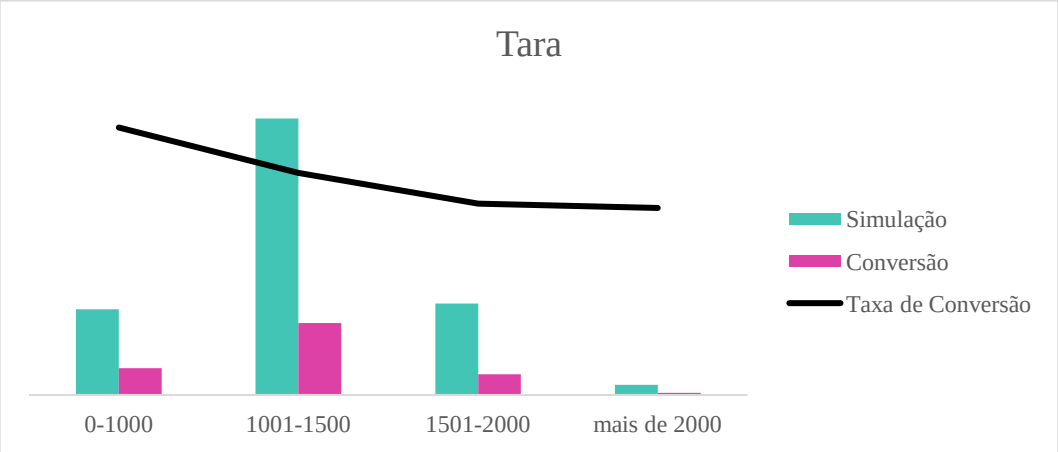

Gráfico 18 - Taxa de Conversão da Variável Tara do Veículo

- Potência do Veículo

O gráfico 19 mostra-nos a variação da taxa de conversão automóvel pela potência do veículo seguro, a maioria da carteira concentra-se em duas bandas: entre os 51 e os 100 e entre os 101 e os 150. Podemos observar que é nos valores de potência dos veículos mais baixa que se encontram as taxas de conversão mais elevadas.

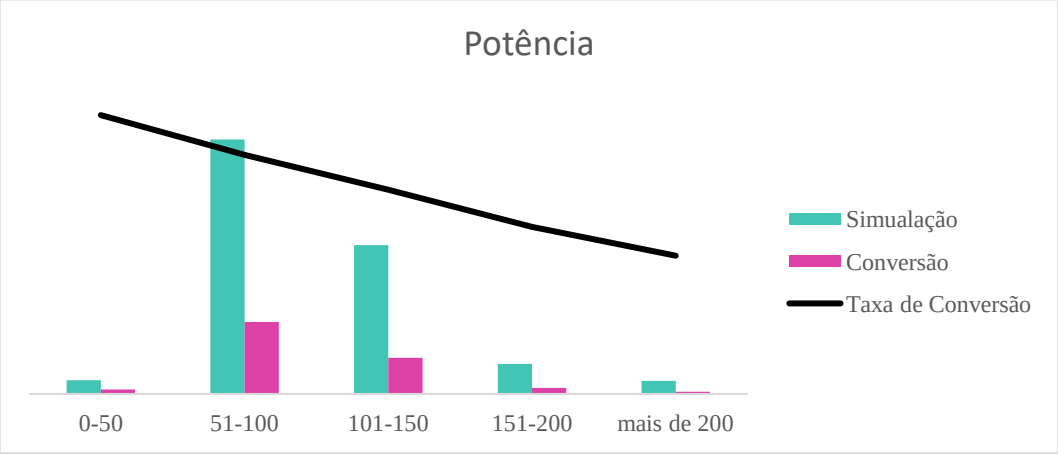

Gráfico 19 - Taxa de Conversão da Variável Potência do Veículo

- Peso Bruto do Veículo

O gráfico 20 apresenta a variação da taxa de conversão automóvel pelo peso bruto do veículo seguro, podemos observar que é entre os 1501 e os 2000 que estão registadas a maior parte das simulações e por conseguinte das conversões.

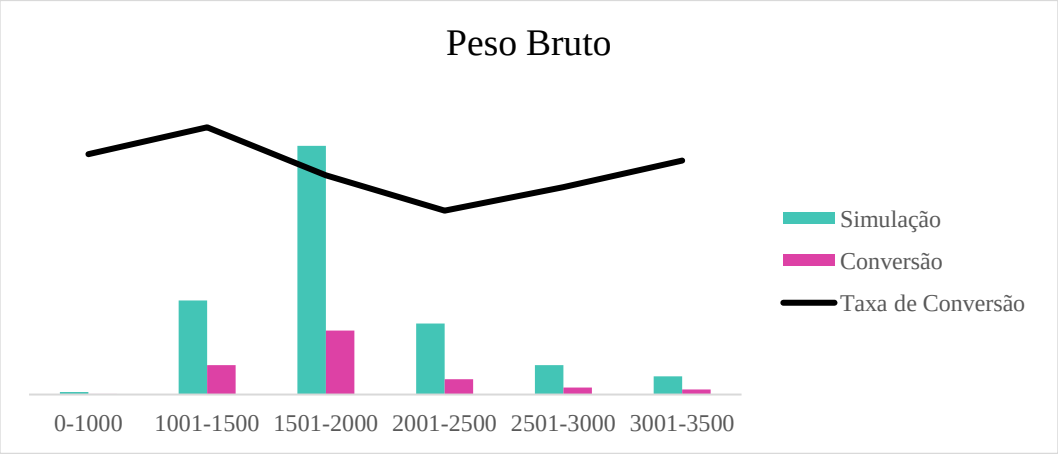

Gráfico 20 - Taxa de Conversão da Variável Peso Bruto do Veículo

- Marca do Veículo

O gráfico seguinte apresenta a variação da taxa de conversão automóvel por marca do veículo seguro, podemos verificar que as marcas mais presentes na carteira são a *Renaut* e a Opel, no entanto as marcas que apresentam uma taxa de conversão mais elevada são a Ford e Fiat.

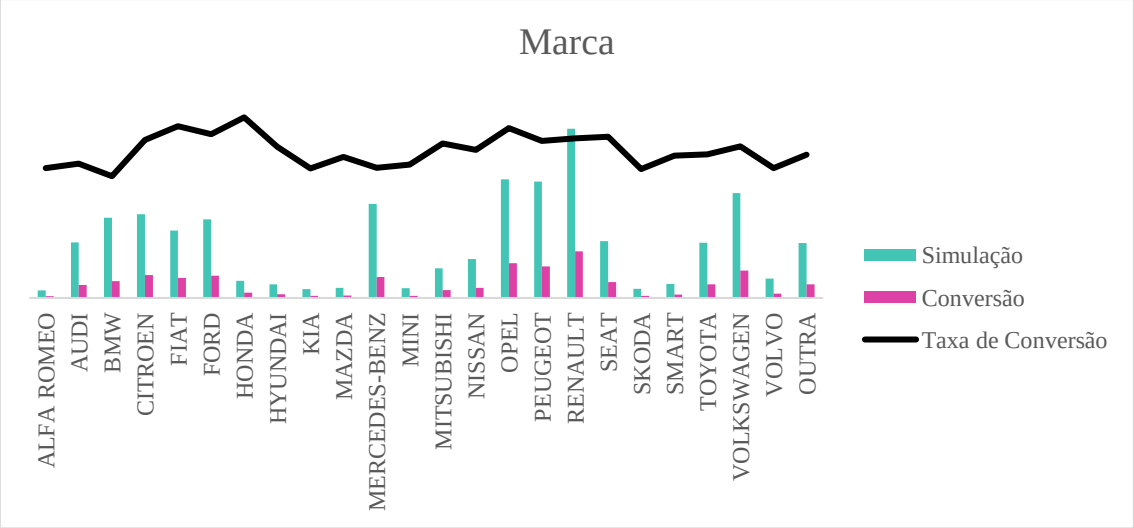

Gráfico 21 - Taxa de Conversão da Variável Marca do Veículo

- Combustível do Veículo

No seguinte gráfico podemos observar a variação da taxa de conversão automóvel pelo combustível do veículo seguro, é em veículos a gasolina e gasóleo que se podem observar conversões mais elevadas.

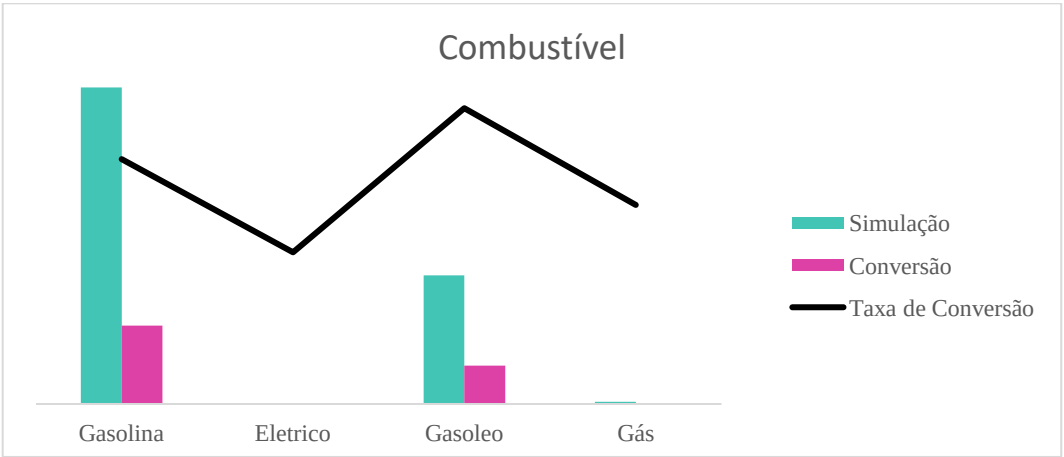

Gráfico 22 - Taxa de conversão da variável Combustível do Veículo

- Caixa de Velocidades do Veículo

No gráfico 23 podemos observar a variação da taxa de conversão automóvel pela caixa de velocidades do veículo seguro. É na caixa manual que está contida a maior parte da carteira.

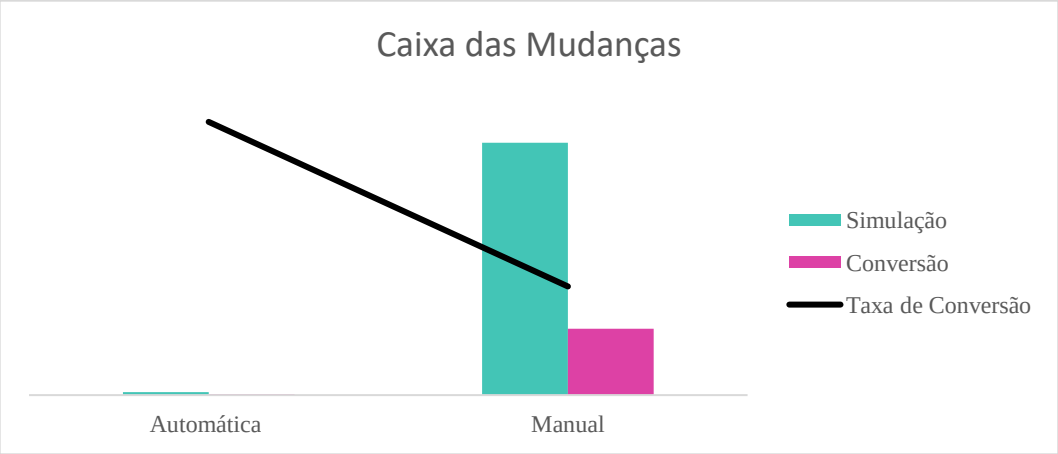

Gráfico 23- Taxa de Conversão da Variável Caixa de Velocidades do Veículo

- Número de Portas do Veículo

O gráfico seguinte apresenta a taxa de conversão automóvel por o número de portas do veículo seguro, como podemos observar a taxa de conversão encontra-se mais elevada nos veículos que possuem 3 portas, sendo os veículos com 5 portas os que apresentam um maior peso na carteira de negócio.

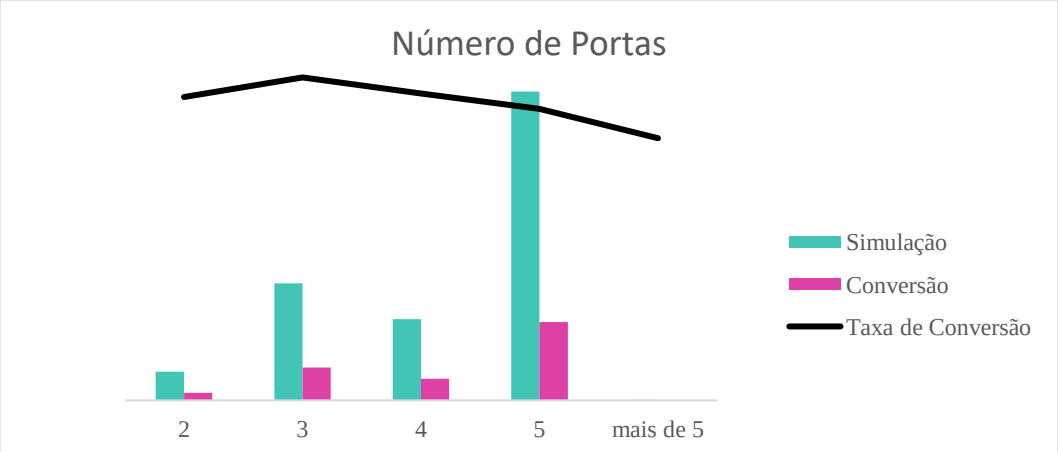

Gráfico 24 - Taxa de Conversão da Variável Número de Portas do Veículo

- Peso Potência do Veículo

O peso potência de um veículo seguro corresponde ao rácio entre o valor da tara e o valor da potência. Através do gráfico 25 podemos observar a variação da taxa de conversão automóvel pelo peso potência do veículo seguro. É nos valores mais reduzidos desta variável que podemos observar uma taxa de conversão automóvel mais elevada.

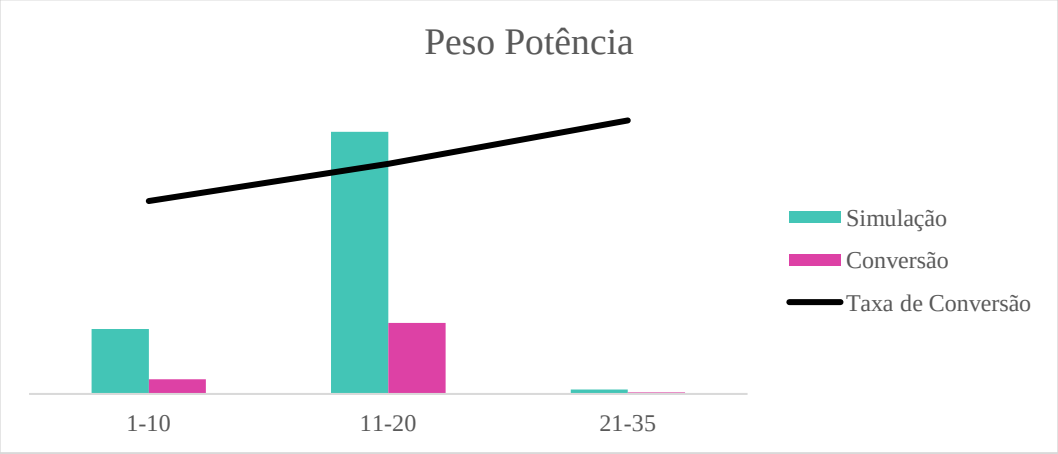

Gráfico 25- Taxa de Conversão da Variável Peso Potência do Veículo

- Controlo de Travagem

No gráfico seguinte podemos observar a variação da taxa de conversão automóvel por controlo de travagem possuído por o veículo seguro. São os veículos que possuem controlo de travagem os que predominam na carteira.

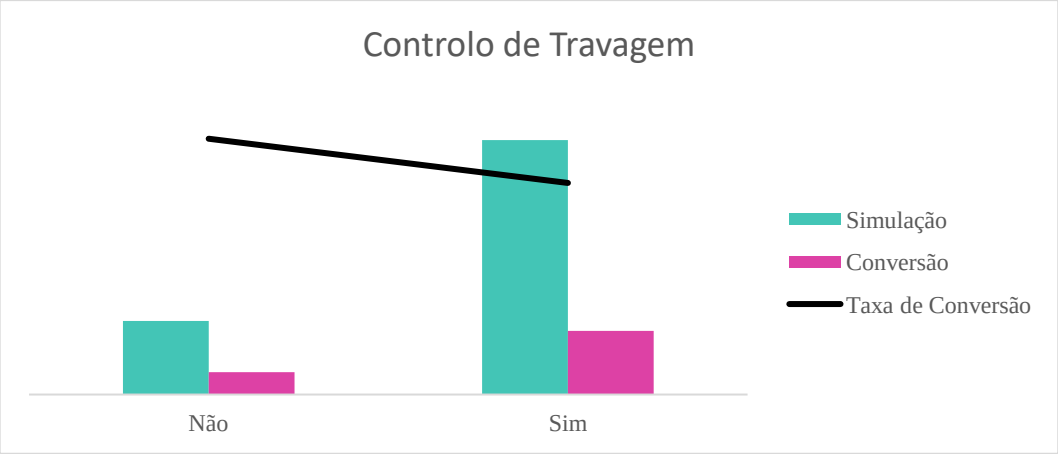

Gráfico 26 - Taxa de Conversão da Variável Controlo de Travagem

- Melhoria da Visibilidade

O gráfico 27 representa a variação da taxa de conversão automóvel por a melhoria de visibilidade, funcionalidade extra que o veículo seguro pode ou não possuir. A maioria dos automóveis seguros que constituem a carteira possuem esta funcionalidade, no entanto a taxa de conversão encontra-se mais elevada nos que não possuem.

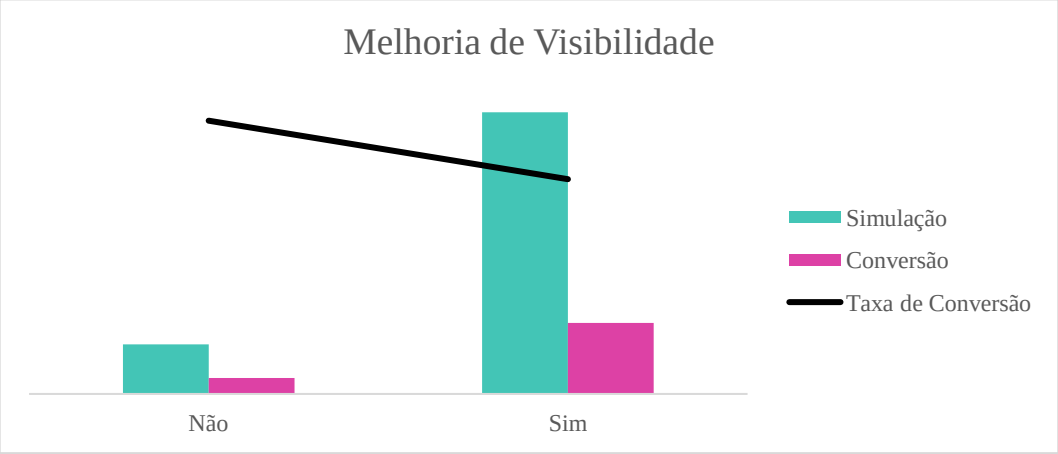

Gráfico 27 - Taxa de Conversão da Variável Melhoria de Visibilidade

- Controlo de Condução

O gráfico 28 representa a variação da taxa de conversão automóvel pelo extra controlo de condução do veículo seguro, onde podemos constatar que não existe uma diferença significativa entre o facto do veículo possuir ou não este extra, no entanto a taxa de conversão é mais elevada nos veículos que não possuem controlo de condução.

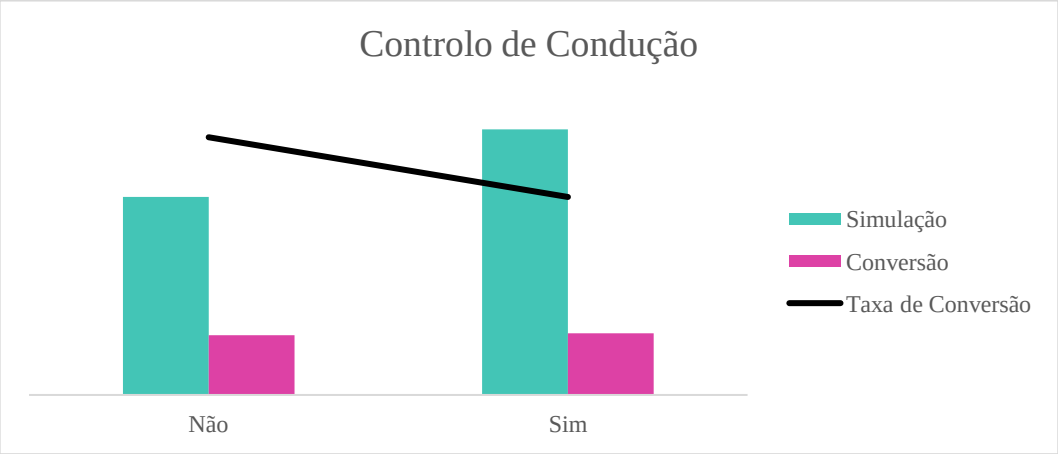

Gráfico 28 - Taxa de Conversão da Variável Controlo de Condução

- Alarme de Segurança

No gráfico 29 está representada a variação da taxa de conversão automóvel pelo alarme de segurança, extra que um veículo pode conter ou não, não existem diferenças significativas entre os dois grupos, no entanto a taxa de conversão é mais elevada para veículos que não possuam este extra.

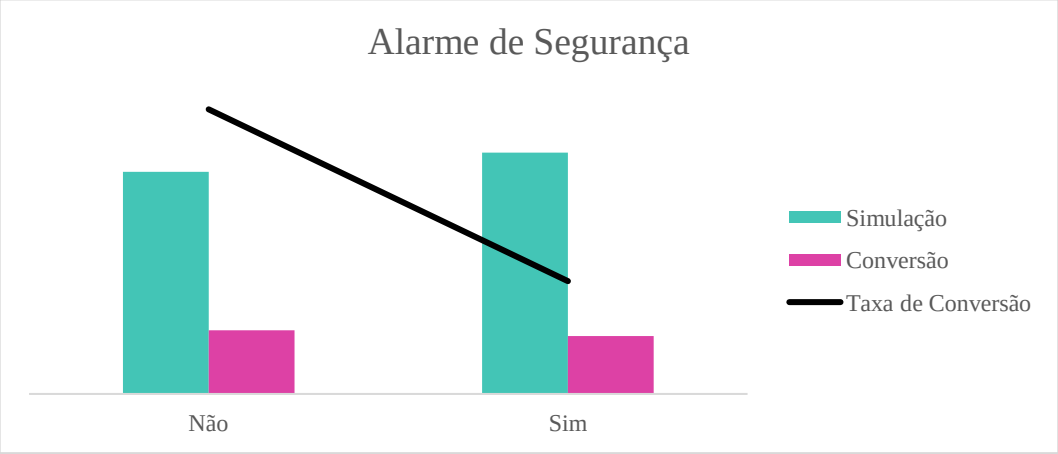

Gráfico 29 - Taxa de Conversão da Variável Alarme de Segurança

- **Características da Apólice**
  - Bónus *Malus*

Quando analisamos a variação da taxa de conversão automóvel por bónus *malus*, podemos verificar que é nos níveis mais elevados que se encontram as taxas de conversão mais elevadas. É no nível 7 que verificamos um maior número de simulações, pois este é o nível inicial para quem nunca tenha possuído um seguro.

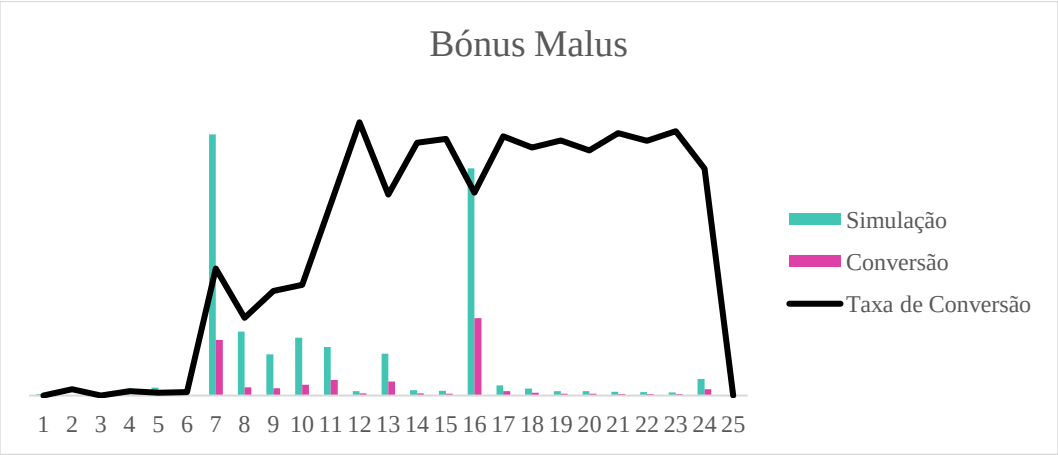

Gráfico 30 - Taxa de Conversão da Variável Bónus *Malus*

- Pack

No seguinte gráfico podemos observar a variação da taxa de conversão automóvel por Pack, podemos observar que é o Pack 1 que possui uma taxa de conversão mais elevada, o que seria de esperar por parte da companhia, uma vez que este é o Pack mais simples e consequentemente que oferece um prémio mais reduzido.

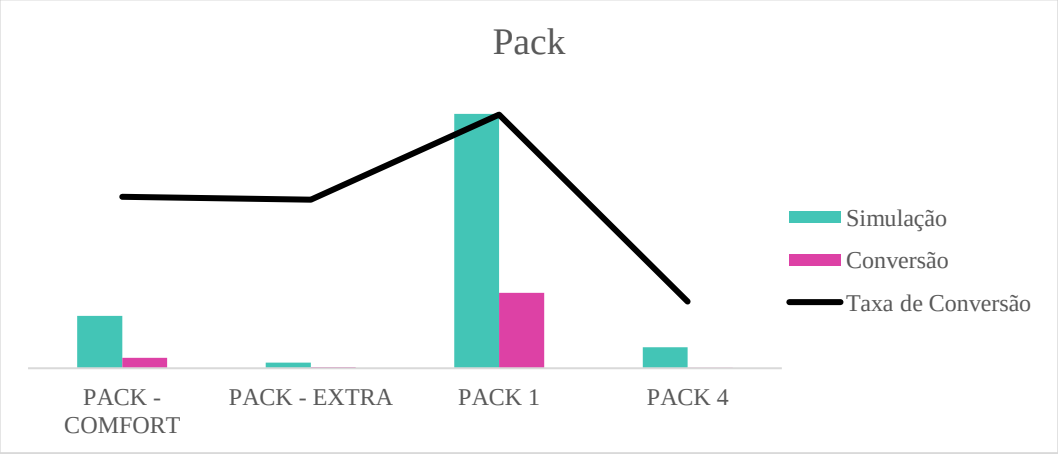

Gráfico 31 - Taxa de Conversão da Variável Pack

- Cobrança Bancária

O gráfico seguinte mostra-nos a variação da taxa de conversão automóvel pela forma de pagamento escolhida pelo cliente, que pode ser bancária ou não, podemos observar que são as transferências bancárias que apresentam valores mais reduzidos de taxa de conversão o que vai um pouco contra as expectativas, uma vez que se um cliente optar por transferência bancária obtém um desconto no prémio.

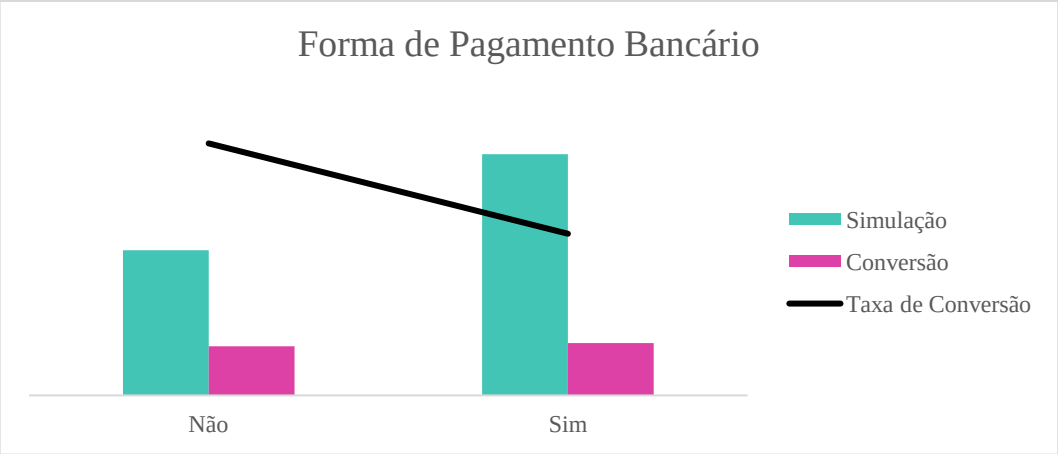

Gráfico 32 - Taxa de Conversão da Variável Forma de Pagamento Bancário

- **Variação do Prémio**
    - Desconto Comercial

O gráfico apresenta-nos a variação da taxa de conversão automóvel por desconto comercial, e como era de prever os clientes que possuem desconto comercial são os que apresentam taxas de conversão mais elevadas, pois o prémio neste caso será mais reduzido.

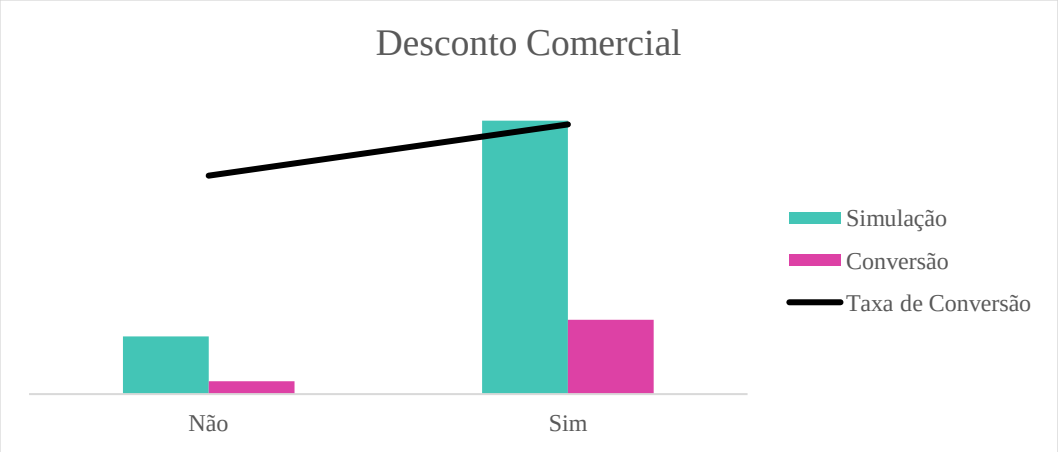

Gráfico 33 - Taxa de Conversão da Variável Desconto Comercial

- Prémio Comercial

O gráfico 34 mostra-nos a variação da taxa de conversão automóvel por prémio comercial, como seria de esperar são os prémios mais baixos que apresentam taxas de conversão mais elevadas.

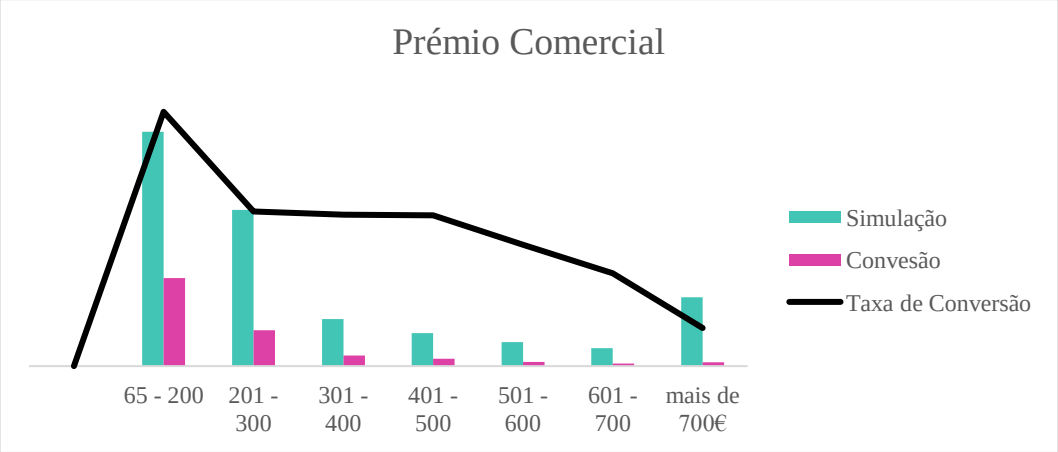

Gráfico 34 - Taxa de Conversão da Variável Prémio Comercial

4.4 Modelação

A base de dados foi criada recorrendo ao *software* SAS Guide, este é uma aplicação do SAS para fazer exploração de dados, permitindo de forma integrada aceder e analisar dados provenientes de diferentes fontes. Para o processo de modelação foi utilizado o *software SAS Miner*, um pacote do SAS para Data *Mining*. O *SAS Miner* contém um conjunto de tarefas de análise que podem ser combinadas de modo a criar e comparar múltiplos modelos. Para além destas funcionalidades, existem tarefas para a preparação dos dados, nomeadamente para deteção de *outliers*, transformação de variáveis, amostragem e partição dos dados em conjuntos de treino, validação e teste (SAS).

Após a criação da base de dados, esta foi dividida de forma aleatória: 40% para treino, 30% para validação e 30% para teste. Esta divisão tem como objetivo verificar se o modelo se encontra bem ajustado aos dados.

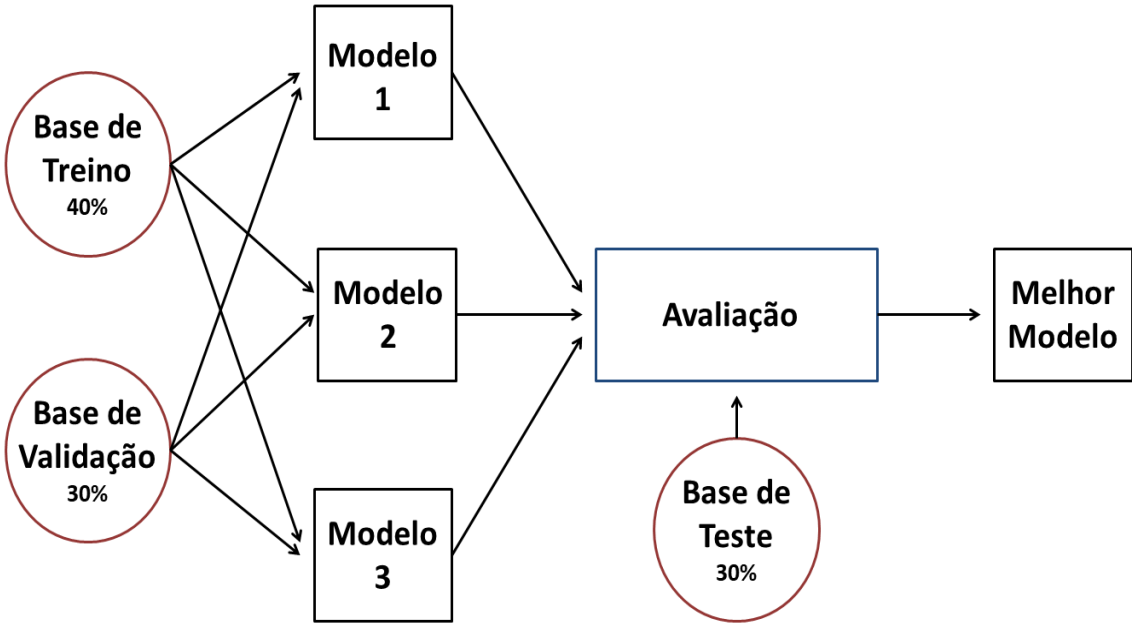

Figura 4 - Representação Gráfica do Processo de Desenvolvimento de Modelos Preditivos

4.4.1 Correlação

A análise da correlação das variáveis é um dos métodos diretos de redução de variáveis. É importante ter em atenção a correlação entre variáveis, pois não deverão ser incluídas variáveis fortemente correlacionadas no modelo, pois estas poderão afetar a qualidade do modelo.

Sempre que um par de variáveis correlacionadas é encontrado, deveremos escolher: a variável que parece ser mais relevante para o estudo em questão ou a variável que apresentar uma melhor qualidade nos dados.

Para a avaliação da correlação entre as variáveis foi utilizado o coeficiente de *V-Cramer*, que corresponde a uma medida de associação entre duas variáveis medidas numa escala categórica. Foi utilizado um limite de 0,7. Na tabela seguinte, podemos visualizar a matriz de correlações de *V-Cramer*.

Variável I	Variável II	V-Cramer
Idade do Condutor	Idade da Carta	1.00
Número da Tara	Peso Bruto	0.89
Peso Potência	Potência do Veículo	0.71
Cilindrada do Veículo	Número da Tara	0.73
Zona DP	Zona RC	0.82

Tabela 4 - Correlações entre as Variáveis, Coeficiente de *V-Cramer*

As variáveis assinaladas com a cor cinzenta foram retiradas do modelo, uma vez que as variáveis descritas na primeira coluna foram consideradas mais importantes para o negócio.

4.4.2 Árvores de Decisão

As variáveis dos nós da árvore de decisão foram selecionadas tendo em conta a importância relativa, ou seja, são selecionadas as variáveis que registam valores de significância mais elevados.

Foram construídos dois modelos distintos, recorrendo às árvores de decisão, um com todas as variáveis e outro sem a variável prémio comercial. Este último modelo foi criado por curiosidade, para podermos ver qual o comportamento das variáveis sem o impacto da variável do prémio comercial, pois esta é a única variável que apenas depende da companhia.

Uma vez que as árvores de decisão são extensas serão interpretados apenas os três primeiros nós de cada uma das duas árvores. Como já foi referido anteriormente, para manter a confidencialidade dos dados da empresa não poderão ser apresentadas as taxas de conversão para cada variável. No entanto, para que a explicação se torne mais clara serão atribuídas taxas de conversão nos nós da árvore de decisão exemplificativas.

4.4.2.1 Árvore de Decisão com Todas as Variáveis

Inicialmente foi construída uma **árvore de decisão com todas as variáveis** que entraram no modelo.

A tabela seguinte apresenta-nos as variáveis que contemplaram a árvore por ordem de importância para o modelo segundo o índice de Gini.

Variáveis	Importância
Bónus Malus	1
Prémio Comercial	0.8802
Transferência Bancária	0.5235
Canal Rede	0.5123
Pack	0.4898
Idade do Condutor	0.4266
Cliente Novo	0.3910
Concelho	0.3801
Idade do Veículo	0.1120
Zona RC	0.1025
Desconto Comercial	0,0682
Combustível	0.0616

Tabela 5 – Importância das Variáveis da Árvore de Decisão que Contém Todas as Variáveis

De seguida serão apresentados os 3 primeiros nós da árvore de decisão, tendo em consideração que o caminho escolhido será sempre onde a percentagem de conversão se encontra mais elevada.

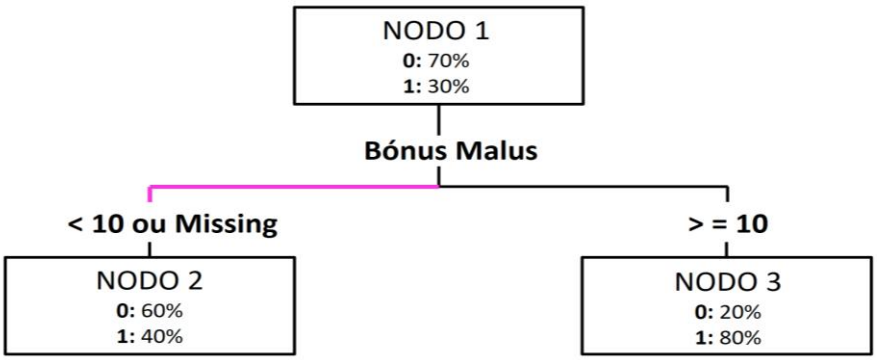

Figura 5 – 1º Ramo da Árvore de Decisão que Contém Todas as Variáveis

O nó de nível superior da árvore de decisão mostra que, de todas as simulações realizadas pelos clientes, onde mais de metade das simulações presentes na amostra não obtiveram conversão. Sob este nó, a variável que melhor caracteriza o modelo é o bónus *malus*, pois é a que apresenta uma maior significância. A linha mais grossa que se apresenta a colorido indica o lado para onde vai o maior número de observações. É para valores de bónus *malus* superiores a 10 que

a taxa de conversão é mais elevada e por isso a árvore de decisão continuará a ser avaliada seguindo pelo nodo 3.

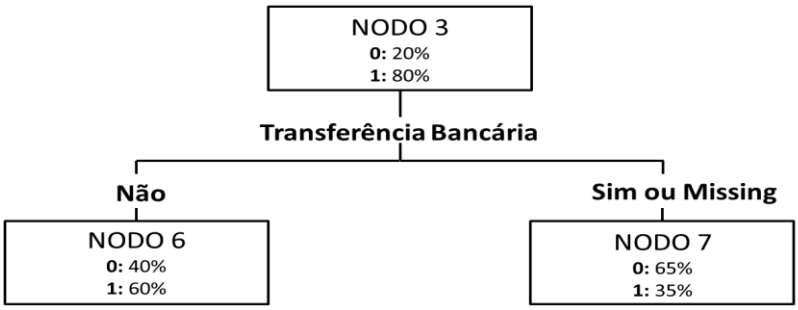

Figura 6 – 2º Ramo da Árvore de Decisão que Contém Todas as Variáveis

Relativamente ao 2º ramo da árvore de decisão, partindo então do nodo 3 como foi decidido anteriormente, podemos constatar que a variável que apresenta uma maior importância neste passo é a variável que diz respeito á forma de pagamento, ou seja, a transferência bancária. No entanto, é a transferência bancaria negativa (clientes que não utilizam transferência bancária) que apresenta uma taxa de conversão mais elevada, segue-se então pelo nodo 6.

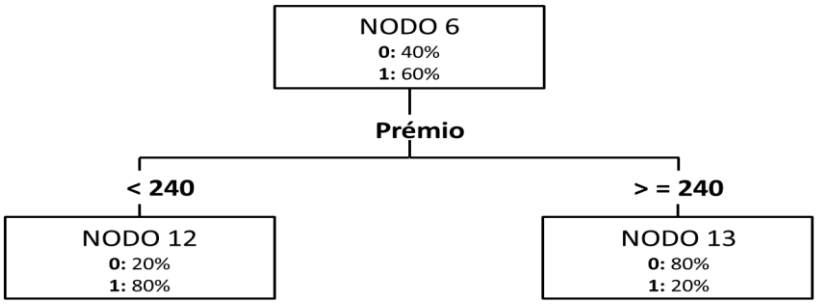

Figura 7 – 3º Ramo da Árvore de Decisão que Contém Todas as Variáveis

No 3º ramo da árvore de decisão que contém todas as variáveis podemos observar que a variável qua apresenta uma importância mais elevada diz respeito ao prémio comercial, e é no nodo 12 que verificamos uma taxa de conversão mais elevada, ou seja, em prémios comerciais inferiores a 240 euros.

4.4.2.1.1 R-Quadrado

A Soma dos Quadrados Total (**SQ_{TOT}**) representa a medida de variação da média, ou seja, a dispersão de cada individuo relativamente à sua média total. O valor de SQ_{TOT} é dado pela seguinte expressão:

$$SQ_{TOT} = SQ_E + SQ_{Trat} \quad \leftrightarrow \quad \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \bar{Y}\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \bar{Y}_j\right)^2 + \sum_{j=1}^{k} n_j\left(\bar{Y}_j - \bar{Y}\right)^2$$

em que, SQ_E corresponde a variação dada pela amostragem aleatória e o SQ_{Trat} corresponde à variação dada pela diferença entre os tratamentos.

O Erro Quadrático Médio (**EQM**) mede a média da diferença entre o valor do estimador e do parâmetro ao quadrado, sendo dado pela seguinte expressão:

$$EQM = \frac{SQ_E}{N - k}$$

em que N corresponde ao número de observações e k corresponde ao número de graus de liberdade. (Gomes, 2016)

As variáveis selecionadas com base no R-Quadrado para a árvore de decisão que contempla todas as variáveis são as seguintes:

Variável	gl	R^2	F	SQ_{TOT}	EQM
Bónus Malus	7	0.043466	1009.93	1308.75	0.1851
Pack	2	0.025629	2141.56	771.682777	0.1801
Cliente Novo	1	0.016592	2823.07	499.566133	0.1770
Canal Rede	6	0.011286	324.05	339.825787	0.1748
Transferência Bancária	1	0.006747	1171.09	203.156092	0.1735
Zona RC	11	0.005361	85.09	161.418618	0.1724
Idade do Condutor	1	0.004308	755.82	129.711987	0.1716
Prémio Comercial	1	0.000142	24.84	4.262680	0.1612
Idade do Veículo	1	0.000105	18.38	3.154050	0.1508

Tabela 6 – Variáveis Selecionadas Segundo o R-Quadrado na Árvore de Decisão que Contém Todas as Variáveis

4.4.2.1.2 Matriz de Confusão

A matriz de confusão da árvore de decisão que contém todas as variáveis apresentadas, compara a classificação real das conversões efetuadas pelos clientes com as conversões previstas pelo modelo. Após serem testados alguns pontos de corte decidimos utilizar um ponto de corte de 60%.

	Previstos		
Observados	Verdadeiro	Falso	Total
Verdadeiro	5330	3713	9043
Falso	35489	111050	146539
Total	40819	114763	155582

Tabela 7 – Matriz de Confusão da Árvore de Decisão que Contém Todas as Variáveis

Pela matriz de confusão anterior podemos calcular a precisão do modelo, sendo que o modelo dado pela árvore de decisão que contém todas as variáveis é de 74.80%.

4.4.2.2 Árvore de Decisão Sem a Variável Prémio Comercial

A fim de compreender se a variável prémio comercial causaria um grande impacto no modelo, foi decidido criar um modelo sem esta variável. A importância das variáveis que contemplaram a árvore de decisão sem a variável prémio comercial segundo o índice de *Gini*, na tabela seguinte:

Variáveis	Importância
Bónus *Malus*	1
Pack	0.8016
Canal Rede	0.6183
Transferência Bancária	0.5962
Concelho	0.4713
Cliente Novo	0.4057
Idade do Condutor	0.3980
Zona RC	0.2576
Tipo de Cliente	0.1935
Combustível	0.1391
Cilindrada	0.0429

Tabela 8 - Importância das Variáveis da Árvore de Decisão Sem a Variável Prémio Comercial

Como podemos observar pela tabela anterior e como tinha sido observado na árvore anterior a variável bónus *malus* é a que apresenta uma maior importância para o modelo.

Serão também avaliados os 3 primeiros ramos desta nova árvore de decisão.

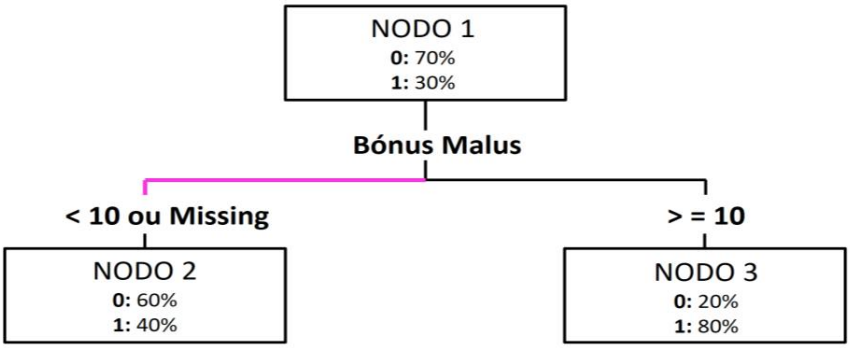

Figura 8 – 1º Ramo da Árvore de Decisão Sem a Variável Prémio Comercial

Mais uma vez podemos constatar que a variável bónus *malus* continua a ser a mais relevante e relativamente a árvore de decisão analisada anteriormente o primeiro ramo não sofre qualquer alteração. Observamos novamente que as taxas de conversão mais elevadas se encontram para valores de bons *malus* superiores a 10 (nodo 3).

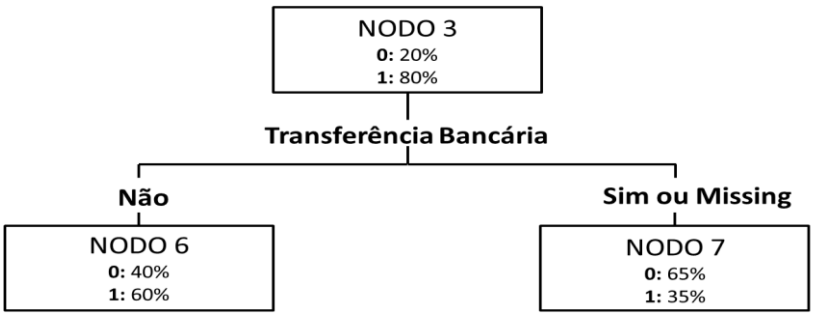

Figura 9 - 2º Ramo da Árvore de Decisão Sem a Variável Prémio Comercial

Neste 2º ramo na nova árvore de decisão também não registamos qualquer diferença relativamente a árvore de decisão anterior. Mais uma vez podemos observar que são os clientes que não pagam à companhia recorrendo a transferência bancaria que apesentam as taxas de conversão mais elevadas e portanto seguimos a nossa análise pelo nodo 6 da árvore de decisão.

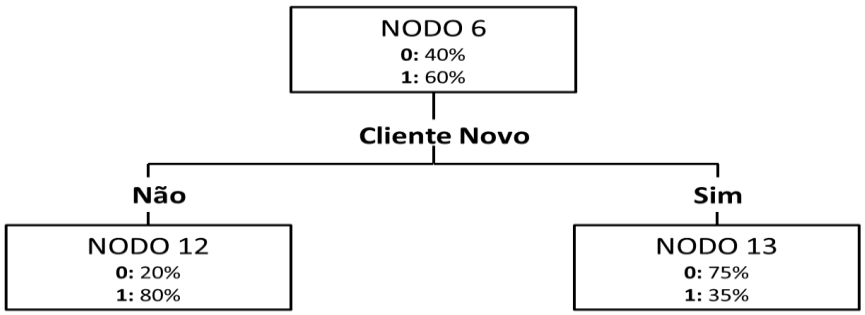

Figura 10 – 3º Ramo da Árvore de Decisão Sem a Variável Prémio Comercial

Relativamente ao 3º ramo, já podemos observar diferenças relativamente à árvore de decisão analisada anteriormente. A variável que neste caso apresenta uma maior importância, é a variável cliente novo. Sendo os clientes que já tiveram algum vínculo com a companhia aqueles que mais convertem (nodo 12).

4.4.2.2.1 R-Quadrado

As variáveis selecionadas com base no R-Quadrado para a árvore de decisão sem a variável prémio comercial são as seguintes:

Variável	gl	R^2	F	SQTOT	EQM
Bónus Malus	7	0.043466	1009.93	1308.75	0.1851
Pack	2	0.025629	2141.56	771.682777	0.1801
Cliente Novo	1	0.016592	2823.07	499.566133	0.1770
Canal Rede	6	0.011286	324.05	339.825787	0.1748
Transferência Bancária	1	0.006747	1171.09	203.156092	0.1735
Idade do Veículo	1	0.000105	18.38	3.154050	0.1615

Tabela 9 – Variáveis Selecionadas Segundo o R-Quadrado na Árvore de Decisão Sem a Variável Prémio Comercial

4.4.2.2.2 Matriz de Confusão

A matriz de confusão da árvore de decisão sem a variável prémio comercial, compara a classificação real das conversões efetuadas pelos clientes com as conversões previstas pelo modelo. Após serem testados alguns pontos de corte decidimos utilizar um ponto de corte de 60%.

	Previstos		
Observados	Verdadeiro	Falso	Total
Verdadeiro	5419	3964	9383
Falso	35400	110799	146199
Total	40819	114763	155582

Tabela 10 – Matriz de Confusão da Árvore de Decisão Sem a Variável Prémio Comercial

Pela matriz de confusão anterior podemos calcular a precisão do modelo, sendo que o modelo dado pela árvore de decisão que contém todas as variáveis é de 74.70%.

Concluímos então, embora que seja por uma diferença bastante reduzida, a árvore de decisão que contém todas as variáveis apresenta uma precisão mais elevada.

4.4.3 Regressão Logística

4.4.3.1 Método de Seleção *Forward*

Tendo em conta toda a amostra de dados é realizada uma análise de regressão logística recorrendo aos três métodos de seleção de variáveis descritos anteriormente.

Será utilizando um nível de significância de 10% para todos os métodos se seleção de variáveis.

Através do método **Forward** obtivemos a seguinte seleção de variáveis:

Variável	gl	Estatística de Wald	Pr > ChiSq
Canal Rede	24	1649.6690	<.0001
Concelho	293	2255.7326	<.0001
Cliente Novo	1	2497.8794	<.0001
Bónus Malus	24	3269.7794	<.0001
Combustível	5	231.1768	<.0001
Marca	68	255.1538	<.0001
Desconto Comercial	1	80.3912	<.0001
Idade do Condutor	1	876.0691	<.0001
Transferência Bancária	1	1078.7477	<.0001
Controlo de Travagem	1	68.4787	<.0001
Urbano/Rural	2	173.9778	<.0001

Pack	4	570.2942	<.0001
Prémio Comercial	1	5.5991	0,018
Tipo de Cliente	1	587.8672	<.0001

Tabela 11 – Variáveis Selecionadas Segundo o Método *Forward* na Regressão Logística

4.4.3.2 Método de Seleção *Stepwise*

Através do método **Stepwise** obtivemos a seguinte seleção de variáveis:

Variável	gl	Estatística de Wald	Pr > ChiSq
Canal Rede	24	1649.6690	<.0001
Concelho	293	2255.7326	<.0001
Cliente Novo	1	2497.8794	<.0001
Bónus Malus	24	3269.7794	<.0001
Combustível	5	231.1768	<.0001
Marca	68	255.1538	<.0001
Desconto Comercial	1	80.3912	<.0001
Idade do Condutor	1	876.0691	<.0001
Transferência Bancária	1	1078.7477	<.0001
Controlo de Travagem	1	68.4787	<.0001
Urbano/Rural	2	173.9778	<.0001
Pack	4	570.2942	<.0001
Prémio Comercial	1	5.5991	0,018
Tipo de Cliente	1	587.8672	<.0001

Tabela 12 - Variáveis Selecionadas Segundo o Método *Stepwise* na Regressão Logística

4.4.3.3 Método de Seleção *Backward*

Através do método **Backward** obtivemos a seguinte seleção de variáveis:

Variável	gl	Estatística de Wald	Pr > ChiSq
Canal Rede	24	1616.2573	<.0001
Concelho	293	1879.7919	<.0001
Cliente Novo	1	2396.6284	<.0001
Bónus Malus	24	3241.4928	<.0001
Categoria	12	72.2287	<.0001
Combustível	5	53.2194	<.0001
Marca	68	231.6757	<.0001
Desconto Comercial	1	74.0557	<.0001
Idade do Condutor	1	698.7664	<.0001
Transferência Bancária	1	955.0248	<.0001
Portas	6	26.8353	0.0002
Controlo Travagem	1	26.2859	<.0001
Melhoria de Visibilidade	1	37.1308	<.0001
Urbano/Rural	2	182.7746	<.0001
Idade do Veículo	1	23.2836	<.0001
Cilindrada	1	45.3330	<.0001
Pack	4	575.6334	<.0001
Tipo de Cliente	1	584.6798	<.0001

Tabela 13 - Variáveis Selecionadas Segundo o Método *Backward* na Regressão Logística

4.5.4 Comparação dos Métodos

Os métodos de seleção de variáveis dos modelos de regressão logística apresentados anteriormente foram comparados recorrendo aos critérios AIC e AUC.

	Forward	Backward	Stepwise
AIC	**154511.13**	154512.70	154576.19
AUC	0.75	0.75	0.75

Tabela 14 – AIC e AUC dos três Métodos de Seleção de Variáveis

Como podemos verificar pela tabela acima, para os três métodos o valor do AUC não se altera, no entanto no que diz respeito ao critério AIC, é o método *Forward* que apresenta o menor valor, sendo assim este o método escolhido.

4.6 Avaliação da Árvore de Decisão versus a Regressão Logística

Como já foi referido anteriormente, da comparação entre os dois modelos relativos às árvores de decisão a que apresentava uma maior precisão nos resultados, e por isso das duas árvores de decisão a considerada mais relevante para o estudo foi a árvore de decisão que contém todas as variáveis que contemplam a base de dados. Relativamente á regressão logística, após terem sido avaliados os três métodos de seleção de variáveis concluímos que o modelo selecionado pelo método *Forward* era o que melhor representava os dados.

Posto isto, com o intuito de perceber qual o método que apresenta o modelo que melhor se adequa aos dados, será feita uma avaliação a fim de compreender qual o melhor modelo, o representado através da árvore de decisão ou o que teve por base a regressão logística.

Através do gráfico seguinte, onde podemos observar as curvas de ROC dos dois modelos que serão comparados. Num bom modelo a curva de ROC deve crescer rapidamente para 1 afastando-se da diagonal (representada a verde no gráfico). Concluímos assim que o modelo de regressão logística selecionado pelo método de seleção *Forward* é o que apresenta a melhor curva de ROC, permitindo-nos concluir que o modelo faz uma boa previsão dos clientes que convertem.

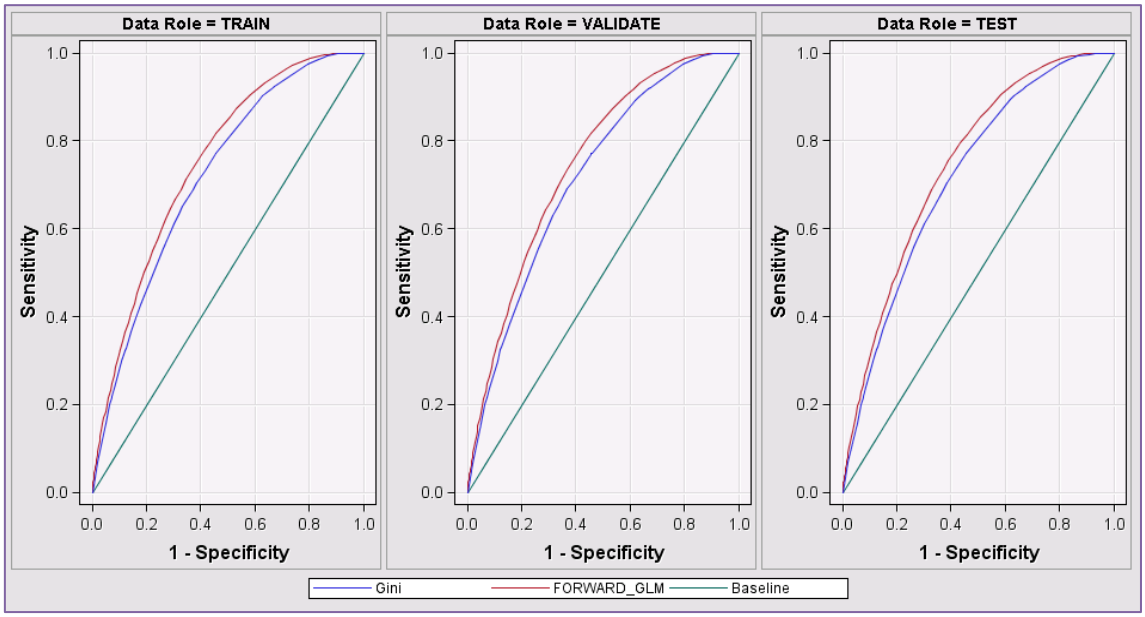

Gráfico 35 – Curvas de ROC da Árvore de Decisão versus o Modelo Selecionado pelo Método *Forward*

Com o intuito de avaliar a capacidade discriminatória do modelo foi calculado o AUC, que corresponde à área abaixo da curva de ROC. Quanto maior for o valor do AUC melhor o modelo explica os dados. Como vimos anteriormente o valor do AUC para o modelo representado pela regressão logística é de 0.74, este valor indica-nos que o modelo tem uma capacidade discriminatória aceitável, não existindo nenhuma evidência que o modelo possa estar desajustado. Como podemos observar através do gráfico anterior, a área representada pela curva correspondente ao modelo representado através da árvore de decisão é inferior. Logo, podemos concluir que o modelo de regressão logística representa melhor os dados.

4.7 Matriz de Confusão

A matriz de confusão do modelo escolhido, compara a classificação real das conversões efetuadas pelos clientes com as conversões previstas pelo modelo. Após serem testados alguns pontos de corte decidimos utilizar um ponto de corte de 60%.

	Previstos		
Observados	Verdadeiro	Falso	Total
Verdadeiro	8801	6925	15726
Falso	32018	107838	139856
Total	40819	114763	155582

Tabela 15 –Matriz de Confusão do Modelo Escolhido

Pela matriz de confusão anterior podemos calcular a precisão do modelo, sendo que o modelo escolhido apresenta aproximadamente 75% dos dados bem colocados. Com este valor de precisão é possível afirmar que as simulações utilizadas no modelo estão bem classificadas.

Capítulo 5: Análise dos Perfis

5.1 Enquadramento Técnico

Na análise e gestão de uma carteira de seguros é necessário ter em atenção alguns indicadores que nos permitem obter, embora que de forma resumida, uma ideia de como a carteira se comporta.

Para analisar a qualidade de um perfil de clientes para a companhia foram avaliados os seguintes indicadores.

5.1.1 Frequência de Sinistralidade

A frequência de sinistralidade é um indicador que nos dá o número esperado de sinistros por período de tempo, normalmente é considerado um ano, num bem seguro ou cobertura.

$$Frequência = \frac{Número\ de\ Sinistros}{Exposição\ ao\ Risco}$$

No mínimo este indicador é zero, quando o número de sinistros é zero quer dizer que não existe valor máximo, logo podemos concluir que quanto menor for este indicador melhor.

Relativamente à exposição, esta dá-nos o tempo, medido em anos, em que o segurador garantiu um bem ou uma cobertura.

$$Exposição\ ao\ Risco = \frac{Número\ de\ dias\ que\ vigora\ a\ apólice}{Número\ de\ dias\ do\ ano}$$

A seguradora apenas está exposta ao risco do conjunto de apólices que se encontram em vigor, ou seja, apenas as apólices que não sofreram anulação. O intervalo de tempo em estudo será então o período desde o início da apólice até a data da sua anulação.

5.1.2 *Loss Ratio*

O Loss Ratio é uma importante ajuda no que diz respeito à análise do lucro da empresa. Quando este valor é 100% significa que recebemos tanto de prémios quanto é gasto com o pagamento dos sinistros, contrariamente, um valor inferior a 100% indica que estão a ser recolhidos mais prémios e assim que a empresa está a produzir lucro.

Este indicador é calculado da seguinte forma:

$$Loss\ Ratio = \frac{Total\ de\ Custos\ com\ Sinistros}{Total\ de\ Prémio\ Adquirido}$$

5.1.3 Custo Médio

O indicador do custo médio permite-nos saber o montante esperado de custos por evento.

$$Custo\ Médio = \frac{Carga}{Número\ de\ Sinistros}$$

$$Carga = Despesas + Indemnizações - Reembolsos + Provisões$$

5.1.4 Prémio Médio

O indicador do prémio médio dá-nos uma média do valor do prémio a ser pago por cliente.

$$Prémio\ Médio = \frac{Prémio\ Anual}{Número\ de\ Apólices}$$

5.2 Perfis

Foram analisados dois perfis de clientes, os que apresentavam uma taxa de conversão mais elevada e contrariamente, os clientes com taxas de conversão mais reduzidas.

Relativamente ao primeiro perfil, os clientes que apresentam taxas de conversão mais elevadas, apresentam as seguintes características: um nível de bónus *malus* superior a 10, ou seja, não têm sinistros há mais de 5 anos, os clientes não optam por transferência bancária no momento do pagamento, são clientes que num espaço de 5 anos já pertenceram à companhia e optam pelo pack que possui apenas responsabilidade civil. Após a análise deste perfil chegamos a conclusão que este é um bom perfil, e portanto, deverão ser desenvolvidas ações para angariar mais clientes com este perfil.

O segundo perfil, referente aos clientes que convertem menos, são clientes com um nível de bónus malus inferior ou igual a 10, o que pode representar taxas de sinistralidade elevadas e clientes que aderem aos packs mais compostos, ou seja com danos próprios. Embora este perfil tenha prémios mais elevados e isso seja uma boa perspetiva de negócio, estes apresentam um *loss ratio* e uma frequência bastante elevada, mostrando que possam vir a simbolizar um mau negócio para a companhia, chegamos então a conclusão que para os clientes que menos convertem não deveremos realizar qualquer ação de modo a captá-los.

Foram ainda avaliados outros perfis de modo a encontrar perfis onde houvesse taxas de conversão mais reduzidas.

Conclusão

A realização do estágio na AGEAS Portugal foi sem dúvida uma mais valia para o início da minha carreira profissional, pois para além de me dar uma visão do mundo profissional proporcionou-me a formação e a descoberta de ferramentas e temas que acabaram por colmatar o os conhecimentos adquiridos tanto na licenciatura como no mestrado. Espero que o projeto que desenvolvi durante este estágio traga benefícios à empresa e que possa ser implementado de modo a que esta aumente a sua rentabilidade ao trazer um maior número de clientes para a companhia.

O objetivo deste projeto consistiu em, delinear perfis de clientes que apresentavam uma maior ou menor propensão à conversão de uma apólice para o seguro automóvel, de uma nova companhia inserida no mercado pela companhia. O conhecimento destes perfis de clientes permite à companhia perceber em quais os clientes que vale a pena investir e contrariamente, os clientes que devem ser evitados, pois estes clientes não trarão uma rentabilidade positiva para a empresa.

Inicialmente, procedeu-se à criação de uma base de dados com todas as variáveis que nos pareceram adequadas para o estudo em questão, e posteriormente cada uma destas variáveis foi analisada de forma a avaliar a consistência dos dados, sendo que as que apresentaram problemas, como um número elevado de missings ou um preenchimento inadequado foram retiradas da base de modo na não prejudicarem o modelo. Após o tratamento das variáveis, procedeu-se a uma análise das correlações para que o modelo não fosse enviesado, neste processo das variáveis que apresentavam uma correlação elevada a escolha recaiu sobre a variável que nos pareceu mais adequada ao negócio. Para a construção do modelo, foram utilizadas duas técnicas distintas, as árvores de decisão e a regressão logística e posteriormente foram comparadas de forma a encontrar o melhor modelo. Finalmente, foi feita uma avaliação do modelo escolhido através da matriz de confusão e do AUC, ambos os métodos traduzem o desempenho do modelo.

De todos os métodos utilizados o que se destacou foi o método de seleção de variáveis forward da regressão logística, que apresenta as seguintes variáveis: Canal Rede, Concelho, Cliente Novo, Bónus *Malus*, Combustível, Marca, Desconto Comercial, Idade do Condutor, Transferência Bancária, Controlo de Travagem, Urbano/Rural, Pack e Tipo de Cliente.

Pelo valor do AUC podemos concluir que o modelo escolhido apresenta uma discriminação favorável, que nos oferece um ajuste a realidade da empresa que pode variar ao longo do tempo, por isso, este modelo deve ser revisto pela empresa com alguma periocidade. Relativamente à matriz de confusão é de salientar que o modelo escolhido apresenta aproximadamente 75% dos valores bem colocados.

Chegamos assim a dois perfis de clientes, os que convertiam mais e contrariamente, os que convertiam menos. Foram discutidas algumas ações para angariar mais clientes que pertençam ao primeiro perfil, de modo a que aumentem as conversões dos consideramos ‘bons’ clientes. No entanto, relativamente ao segundo perfil discutido, os clientes que convertiam menos eram exatamente aqueles clientes que as companhias seguradoras não querem que façam parte da sua carteira e por isso nada será feito de modo a cativa-los para que estes tenham um seguro na companhia.

A criação de modelos estatísticos, como o que foi apresentado anteriormente, tem se tornado cada vez mais importante para as empresas, pois permitem retirar bastante informação que poderá ser usada noutras linhas de negócio.

O modelo deverá ser atualizado pela companhia, todos os anos, uma vez que o mercado segurador está em constante crescimento, devendo ser recalibrado.

Bibliografia

1. Ageas no Mundo, Apresentação da Empresa (Março 2019). Disponível em: https://www.grupoageas.pt/sobre-o-grupo-ageas/a-ageas-no-mundo

2. Ageas Portugal, Apresentação da Empresa (Março 2019). Disponível em: https://www.grupoageas.pt/sobre-o-grupo-ageas/quem-somos

3. Alpuim, T. (2018). *Notas de Apoio à disciplina de Modelos Lineares.*

4. ASF (2015) – Guia de Seguros e Fundos de Pensões (3º edição). Disponível em: https://www.asf.com.pt/NR/rdonlyres/3F64D61A-BCDE-48F3-8C36-244F65CD9CAE/0/GuiadeSeguroseFundosdePens%C3%B5es_2015.pdf

5. Associação entre variáveis (outubro 2019). Disponível em: http://sweet.ua.pt/andreia.hall/TEA/Capcorrel.pdf

6. Breiman, L. & Friedman, J. & Stone, C. J. & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis LTD.

7. Carvalho, H. (2017). *Estudo da Vinculação de um Cliente Particular a um Banco*. Tese de Mestrado, Universidade de Lisboa – Faculdade de Ciências.

8. Chatterjee, S. & Hadi, A. S. (2012). *Regression Analysis by Exemple* (5th Edition). John Wiley & Sons Inc.

9. Esteves, R. (2018). *Notas de Apoio à disciplina de Atividade Seguradora.*

10. Figueiredo, P. (2018). *Notas de apoio à disciplina de Atividade Seguradora.*

11. Freund, R. J., Wilson, W. J., Sa, P. (2006) – *Regression Analysis, Statistical Modeling of a Response Variable* (2nd Edition). Academic Press.

12. Garraio, J. (2015). Modelação da Taxa de Anulação no Seguro Automóvel. Tese de Mestrado, Universidade de Lisboa – Faculdade de Ciências.

13. Ginja, A. (2017). Modelação da Zona Tarifária no Seguro Automóvel. Tese de Mestrado, Universidade de Lisboa – Faculdade de Ciências.

14. Glossário Ageas, Glossário de Seguros (março 2019). Disponível em: https://www.ageas.pt/glossario-de-seguros

15. Gomes, J. (2016) *Apontamentos à disciplina de Estatística Aplicada*.

16. Guedes Vieira, M. (2012). *Introdução aos Seguros*. Vida Económica, Porto.

17. Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2ª Edition). New York: USA: A Wiley-Interscience Publication, John Wiley & Sons Inc.

18. Jong, P. & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data.* New York: Cambridge University Press.

19. Kleinbaum, D. G. & Klein, M. (2002). *Logistic Regression* (3ª Edition). New York: Springer-Verlag New York Inc.

20. Lewis, R. (2000) ―*An introduction to classification and regression tree (CART) analysis*‖, Annual Meeting of the Society for Academy Emergency Medicine, San Francisco, California. USA.

21. Lift Chart (Analysis Services – Data Mining). Disponível em: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=sql-server-2017

22. McHugh, M. (2013) – The Chi-square test of independence, Lessons in Biostatistics. *Biochemia Medica* 2013;23(2):143-149.

23. Mendenhall, W. & Sinich, T. (2011). *A Second Course in Statistics – Regression Analysis* (7ª Edition). Pearson.

24. Moro, S. (2011). *Otimização da Gestão de Contactos Via Técnicas de Business Intelligence: Aplicação na Banca*. Tese de Mestrado, Universidade de Lisboa – ISCTE – Instituto Universitário de Lisboa.

25. Moro, S. & Laureano, R. *Using Data Mining for Bank Direct Marketing: An Application of the Crisp-Dm Methodology*. Disponível em: https://repositorium.sdum.uminho.pt/bitstream/1822/14838/1/MoroCortezLaureano_DMApproach4DirectMKT.pdf

26. Mousinho, G. (2016). *Modelling Renewal Price Elasticity: An Application to the Motor Portfolio of Ocidental.* Tese de Mestrado, Universidade de Lisboa – Instituto Superior de Economia e Gestão.

27. Silva, C. (2018). *Modelação da Elasticidade do Preço na Renovação Automóvel*. Tese de Mestrado, Universidade de Lisboa – Faculdade de Ciências.

28. Turkman, M. A. A. & Silva G. L. (2000). Modelos Lineares Generalizados – *Da Teoria à Prática.*